APPLYING JOHN ROEMER'S THEORY OF REDISTRIBUTIVE JUSTICE AS ALGORITHMIC FAIRNESS

SANDER LEENAARS

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

734641

COMMITTEE

dr. M.W. Klincewicz
dr. M.M. Louwerse

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

January 14, 2022

WORDCOUNT

8679

CODE CAN BE REQUESTED AT

https://github.com/sanderleenaars/algorithmicfairness.git

ACKNOWLEDGMENTS

I would like to thank professor Klincewicz for his supervision during the writing of this thesis. He guided me towards this particular subject, which is a perfect combination of my economical bachelor and current master program. It enabled me to use knowledge gathered in a variety of courses such as machine learning, data ethics & data law. I also like to thank Maria Soares da Eira who was the first to apply a social-economic theory to algorithmic fairness.

APPLYING JOHN ROEMER'S THEORY OF REDISTRIBUTIVE JUSTICE AS ALGORITHMIC FAIRNESS

SANDER LEENAARS

Abstract

Machine learning models can discover patterns that are not immediately apparent. However, they bring a substantial risk to outcome decisions since human bias can influence data collecting, on the one hand, and, on the other hand, the transparency or explainability of why a certain prediction is made, might be lost. To guarantee machine learning models are not (unintendedly) producing discriminatory outcomes numerous mathematical fairness metrics and fairness-enhancing methods have been developed. However, most of the fairness metrics have lost a substantive connection to theories of justice and fairness as it is conceived of in society. To address this gap, this thesis first offers in-depth exploration of the social-economic theory developed by Roemer about redistributive justice. It then proposes a new algorithm to expose individuals with identical relative effort levels to an equal probability of a positive outcome prediction. This machine learning is then tested with the COMPAS dataset to predict the likelihood of a criminal reoffending. Results show that the algorithm based on Roemer's theory increases several fairness metrics values while retaining a robust connection to a normative conception of justice operative in society.

1 INTRODUCTION

Machine learning and automated algorithms play a substantial role across all facets of current society. They are adopted in a wide variety of circumstances from ranking job applicants, predicting income, creating recommender systems, to predicting recidivism. These powerful tools are under continuous improvement with new technologies that achieve ever higher accuracies in both prediction and classification tasks. However, when used incorrectly they can have disastrous societal outcomes. An extensive number of scientific papers have illustrated that machine learning models can generate discriminatory outputs much alike human biases present in

society (Jiang & Nachum, 2020; Najibi, 2020; Zhong, 2018). This created an urgent need for a new methodology that can carefully assess the fitness of training data and actively monitor for fairness and transparency.

Over the last years, a considerable body of research focused on fairness-enhancement algorithms to combat human biases present in a dataset (Corbett-Davies, Pierson, Feller, Goel, & Huq, 2017; Feldman, Friedler, Moeller, Scheidegger, & Venkatasubramanian, 2015; Kamiran & Calders, 2012). This has resulted in several algorithmic fairness metrics, each offering their own understanding of what "fairness" means. In this debate there seems to be no universal agreement on the definition of fairness, even though what fairness amounts given a particular context must be carefully fitted to the needs of the task and the data used.

This thesis explores one particular definition of fairness rooted in social-economic theories, which can be used across a number of contexts, with any machine learning algorithm, and in both predictive and classification paradigms. The linchpin of this approach is a theory of redistributive justice developed and defended by John Roemer (Roemer, 1998). This social-economic theory postulates that income disparities in society due to different levels of personal effort should be allowed. However, Roemer observes, the absolute level of effort that an individual can provide might be restrained by circumstances. Given this, Roemer created a notion of relative effort that exemplifies effort provided after correcting for circumstances of an individual.

To apply this theory of redistributive justice to the machine learning domain, a corrective post-processing algorithm can be developed. This algorithm should assure that classes of people labelled in a dataset are counted as having exerted the same degree of relative effort and receive an equal proportion of positive label predictions whilst minimizing accuracy loss. To achieve this, I first developed a mixed effects logistic regressor that was trained on the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset and used this model to predict the likelihood of a criminal reoffending. The COMPAS dataset is widely used by researchers to contest a racial bias against African Americans.

The new machine learning model is then evaluated against numerous fairness metrics to measure performance and the results answer the following research question:

> *To what extent can John Roemer's social-economic theory of redistributive justice be applied to increase algorithmic fairness in the COMPAS dataset?*

Several auxiliary research questions are answered along the way:

RQ1 *What are the most prevailing techniques constructed in the algorithmic fairness domain?*

RQ2 *What variable(s) in the COMPAS dataset best exemplify the degree of effort provided by an individual to receive a positive outcome?*

RQ3 *To what extent can a method be developed to achieve equal opportunity of receiving a positive label prediction for individuals with the same level of relative effort?*

RQ4 *How does this post-processing model perform compared to other bias mitigation algorithms measured on multiple fairness metrics developed previously in the algorithmic fairness domain?*

2 RELATED WORK

2.1 *Discrimination mitigation algorithms*

According to a survey carried out in 2018 by Groen (2018) suppression is the most widely used pre-processing method by businesses to increase fairness in algorithmic decision-making. This method is known as fairness through blindness and originated in social sciences which concluded that human decision-making should never be based on protected attributes such as race, sex or religion unless this is of vital importance for the person itself. Similarly, machine learning models should not include protected attributes as input. Not surprisingly, suppression is the most adopted method by businesses because it also is the minimum requirement set by law. The GDPR, founded in 2018 by the European Union, prohibits any processing of "sensitive data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data, data concerning health or data concerning a natural person's sex life or sexual orientation".(*General Data Protection Regulation*, 2018, art. 9.). Unfortunately, scientific research indicated that this prominently adopted method is not merely enough to prevent discriminatory decision-making (Barocas & Selbst, 2016; Hajian, Bonchi, & Castillo, 2016). Protected attributes can indirectly affect other features and steer an algorithm towards discrimination (Giannotti & Pedreschi, 2008). One well-known precedent is redlining: discrimination based on a neighbourhood or region an individual lives in. Postal code has repeatedly proven to proxy ethnicity and people from a specific racial background

have suffered from a discriminatory denial of service (Calders & Verwer, 2010).

In an attempt to overcome this problem, scientific literature proposed a variety of bias mitigation method that actively monitor fairness metrics. During this process, three different branches of algorithmic fairness were established: independency, separation and sufficiency. These concepts and their notable differences will be explained in the next three paragraphs. In order to provide a descriptive answer to RQ1, the most widely adopted algorithms, together with their adopted algorithmic fairness metric, are explained. These methods will be recreated and fitted on the COMPAS dataset as well to enable the comparative analysis required to answer RQ4.

2.1.1 *Independency*

Statistical parity difference (SPD), also known as demographic parity or independence is one of the most extensively considered fairness metrics (Corbett-Davies et al., 2017). Contrary to suppression which relies on fairness through blindness, this definition actively monitors if unfairness occurs by comparing the positive label predictions for each group. SPD assumes no dependency between sensitive attributes and outcome. Therefore, no accuracy component is incorporated in its' equation. SPD subtracts the proportion of positive predicted labels of the disadvantaged class from the proportion of predicted positive labels for the advantaged class.

$$SPD = P(\hat{y} = 1|z = 0) - P(\hat{y} = 1|z = 1) \tag{1}$$

An SPD of zero guarantees perfect equality such that both groups have the same exact percentage of positive label predictions. However, values within the boundaries of -0.1 and 0.1 are considered as fair. A relaxed version developed by Zafar, Valera, Gomez Rodriguez, and Gummadi (2017) called the 'p% rule', or disparate impact (DI) is most used in practice.

$$DI = \frac{P(\hat{y} = 1|z = 0)}{P(\hat{y} = 1|z = 1)} \tag{2}$$

This relaxation is regularly linked to the four-fifth rule which assumes disparate impact occurs when predictions of a positive class is less than four-fifths of the most privileged group. This has gained traction because of its legal support founded in 1978. The Equal Employment Opportunity Commission (EEOC) ruled that the selection rate for any race, sex or ethnic group should have at least an 80% selection rate compared to the group with the highest selection rate (*EEOC compliance manual.*, 1978). SPD became the first metric for measuring unintended discrimination (Feldman et al., 2015). Adopting the SPD as a metric would therefore guarantee

legal support and has proven to create a Pareto suboptimal that improves the representation of disadvantageous groups in the labour market (Hu & Chen, 2018).

One popular fairness method optimizing this SPD fairness metric is reweighing (Kamiran & Calders, 2012). Similar to the SPD, the methodology assumes generated predictions should not reveal any conditional dependence on sensitive features. To assure that predictions and sensitive attributes are statistically independent it distributes all instances to four different groups. These groups are divided by privileged or unprivileged group and favourable or unfavourable outcome. Once, every data point is assigned to a group, the favoured group will be assigned a lower weight and the unfavoured group will receive a higher weight. "The weight of an object will be the expected probability to see an instance with its sensitive attribute value and class given independence, divided by its observed probability" (Kamiran & Calders, 2012). Finally, the number of occurrences of every object is multiplied with its' weight to remove all discrimination present in the dataset.

Another popular bias mitigation algorithm is Adversarial debiasing (Zhang, Lemoine, & Mitchell, 2018). This model states that the output from a classifier should not be useful in predicting sensitive attributes. This shows that Adversarial debiasing, similar to SPD, aims for conditional independence between the protected attributes and outcome label predictions. This so-called Generative Adversarial Net (GAN) is a method that simultaneously trains two neural networks with the first model optimizing a loss function and a second model adversing the first model with some discriminative factor. Adversarial debiasing is a specific type of GAN where the adversarial neural network obtains output from the generative neural network and tries to predict the sensitive attributes using only this generated output. The goal of these inter-twining neural networks is to guarantee that no information about the protected attributes is leaked during the training stage. The following process describes best the workflow of Adversarial debiasing:

1 Train the *generative* neural network to predict the target variable

2 Train the *adversarial* neural network to utilize the output from the generative neural network to predict the sensitive attribute

3 Iterate over the process

This Adversarial debiasing technique developed by Zhang et al. (2018) competes in a zero-sum game for the number of epochs and corrects,

at each epoch, by the respective learning rate determined at the cross validation stage. The ultimate goal is to optimize accuracy whilst assuring the adversarial neural network cannot predict the protected category of each instance.

2.1.2 *Separation*

Hardt, Price, and Srebro (2016) proposed a notion called equalized odds, also known as separation. Separation differs from independency as it allows predictions to be dependent on the sensitive attributes but only via the target label. Given the fact that outcome (y) occurred, the proportion of positive label predictions should be equal across groups. Equalized odds thus assumes that the predictions and sensitive attributes are conditional on the target labels.

$$Equalized_{\text{odds}} = P(\hat{y} = 1|z = 0, y = b) - P(\hat{y} = 1|z = 1, y = b), b \in \{0, 1\} \tag{3}$$

This fairness metric evaluates how well a classifier predicts a positive label for any target label across groups. Equalized odds measures the differences in true positive rate (TPR) and false positive rate (FPR) across all classes. The ideal value being 0 but concluded to be fair between -0.1 and 0.1. Since it optimizes across the TPR and FPR, equalized odds presumes that the ground label is true and reliable. Advantageously, the outcome is now compatible to sensitive attributes which might show different underlying distributions (Ghassami, Khodadadian, & Kiyavash, 2018). However, if human bias was incorporated during the data collection, the validity of ground label y becomes questionable. If, for example, individuals with a specific ethnicity obtain more erroneous arrests, and this is not corrected at the data pre-processing stage, individuals from this ethnicity will be assigned more frequently to the disadvantageous outcome. In such cases, optimizing for fairness metrics with an accuracy component potentially amplifies discrimination already present in the dataset (Garg, Villasenor, & Foggo, 2020). It is therefore important to always consider whether the ground labels are reliable.

In binary classification tasks, often only the advantaged outcome is considered. Therefore, a relaxation called equal opportunity difference (EOD) is created (Hardt et al., 2016).

$$EOD = P(\hat{y} = 1|z = 0), y = 1 - P(\hat{y} = 1|z = 1), y = 1 \tag{4}$$

Equality of opportunity only requires non-discrimination across the positive label predictions. True positive rate should now be similar between the protected and unprotected class.

Hardt et al. (2016) created an algorithm based on the EOD. As shown in the equation, techniques enforcing the EOD fairness metric require input about the predictions, target label and sensitive attribute for each instance. Figure 1 graphically illustrates ROC curves of the TPR and FPR for all classes at various thresholds.

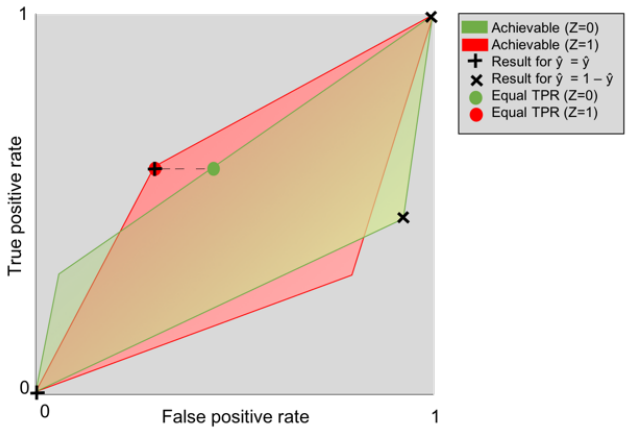

Figure 1: Space of possible predictions for the unprivileged and privileged classes

Equalized odds explores a range of positions where the probability of obtaining a positive label prediction given a positive outcome (TPR) is equal across groups. The optimal solution falls within the overlapping quadrilaterals as shown in figure 1. The green quadrilateral represents the unprivileged class and the red quadrilateral represents the privileged class. The four different dots are constructed by taking the extreme outcomes as well as the TPR and FPR points. The equal opportunity metric is satisfied when the TPR of both classes are equal as illustrated by the red and green dot. The authors turn this into a linear problem such that it searches for which label predictions to alter in order to minimize the loss function but still adhering to the TPR constraint.

2.1.3 *Sufficiency*

Zafar et al. (2017) invented predictive rate parity, also called sufficiency or disparate mistreatment. Contrary to equalized odds, this definition assumes that the target variable Y is independent of the sensitive attributes but is conditional on the predictions. Disparate mistreatment occurs if the misclassification rates vary for groups with different sensitive attributes. This definition combines the false positive rate and false negative rate.

$$FPR = P(\hat{y} \neq y | z = 0, y = 0) - P(\hat{y} \neq y | z = 1, y = 0) \tag{5}$$

$$FNR = P(\hat{y} \neq y | z = 0, y = 1) - P(\hat{y} \neq y | z = 1, y = 1) \tag{6}$$

A FPR and FNR of 1 guarantee equality across the positive label predictions for both groups. However, values falling within 0.8 and 1.2 are considered to be fair. In relation to both the FPR and FNR, two scenarios can be encountered. (i) FPR and FNR have opposite signs. The algorithm has developed a bias to a group with a specific sensitive attribute. It assigns the privileged group disproportionally to the positive outcome (resulting in a high FPR) and assigns it significantly less to the negative outcome (resulting in a low FNR). (ii). FPR and FNR have a uniform sign. The unprivileged group obtains a higher FPR and higher FNR. This scenario might occur when a group with a certain sensitive attribute is harder to classify and therefore performs worse on all misclassification rates.

Similar to equalized odds, disparate mistreatment performs best when the ground truth is reliable and available (Zafar et al., 2017). The FPR and FNR are set as constraints at the in-processing stage while a logistic regression model aims to optimize the original loss function to achieve the highest accuracy possible.

All these methods have proven their incredible capability to remove discrimination in outputted predictions. However, because these methods are incorporated to the decision-making process, it has become increasingly complex and opacity becomes a real concern. The reason to why a certain prediction is made can quickly become incomprehensible when adversarial neural nets and quadrilaterals are introduced. This can become an evident problem because transparency is of vital importance. Both the programmer and the individual who is subject to a certain label prediction, deserve the right to know why a particular decision is made. In order to address this issue and offer such an explanation, the following paragraph will explore John Roemer's social-economic theory of redistributive justice. Exploring and incorporating beliefs rooted in social-economic theories can guarantee that the executed definition of fairness is consonant with society's conception of justice. This social-economic theory will be used to develop an algorithm that is in line with the reasoning of Roemer's theory and therefore contributes in answering the last three research questions.

2.2 *Roemers' social-economic theory of redistributive justice*

Arguably one of the most important contributions of social sciences is the discussion as to which kinds of inequality are ethically justifiable. In 1958 Rawls was the first to form a novel approach to egalitarianism (Rawls,

1958). His theory stated that personal choices and the related personal responsibility should be a crucial determinant in the equality equation. Rawls (1958) aimed to offset differences due to luck but not differences in outcomes caused due personal choices. This new theory replaced equality of outcomes with an equality of opportunity approach (Roemer & Trannoy, 2016). Outcomes are assumed to be combination of luck which individuals cannot influence and actions which are deliberately taken and considered by an individual.

Roemer build-upon this Rawlsian definition in his book: Theories of distributive justice. Roemer (1998) divides a population into a finite set of groups. These divided groups consists of people with the same *circumstance*. Circumstance are occurrences beyond one's control but affect the outcome an individual receives. Possible examples of such circumstances are biological characteristics, the family a person is born in, educational parental levels and more. Roemer argues that an individuals' income is partly caused by circumstances and partly caused due to personal *effort*. Effort is defined as "a choice made by the individual, although that choice may be constrained by circumstances, that affects the outcome of interest".(a case which will be discussed in the next paragraph) (Roemer & Trannoy, 2016). The redistributive justice theory of Roemer regards differences in income between individuals caused by their circumstance (e.g. ethnicity) as ethically unacceptable and income differences due differential levels of personal effort (e.g. working more hours) as ethically acceptable. Roemer's redistributive justice theory therefore fixates on offsetting differences due to circumstances but not on personal effort.

To illustrate his theory, Roemer and Trannoy (2016) define adult wage as outcome. This theory differs from egalitarianism as it allows different wages due to different levels of effort provided by individuals (e.g. an individual that chose to study for a longer period is allowed to earn more). However, the number of years studied might be constrained by certain circumstances. One such circumstance that is addressed, is parental income levels. Higher level of parental incomes can afford higher intuitions leading to the possibility to study for a longer period of time. Roemer argues that an individual should not be 'punished' and receive a lower wage for being born in a family with a low parental income. This examples shows how effort (years studied) is constrained by circumstances (parental income), and wage can therefore not be the result of raw levels of effort. A notion of *relative effort* should be developed to eliminate circumstance from absolute levels of effort.

High parental income (circumstance)	Years of education (effort)	Percentile
Individual A	4	25th
Individual B	5	50th
Individual C	3	0th
Individual D	8	100th
Individual E	7	75th

Table 1: Years of education and their respective percentile for the high parental income group

Low parental income (circumstance)	Years of education (effort)	Percentile
Individual F	5	100th
Individual G	1	0th
Individual H	2	25th
Individual I	4	75th
Individual J	3	50th

Table 2: Years of education and their respective percentile for member of the low parental income group

Table 1 and 2 will be used as an illustration as to how relative effort is formed. Years of schooling is set to be effort e and parental income is set to be circumstance c. Imagine that individual D has obtained 8 years of education and individual F only has 4. A consequentialist would argue that individual D has earned the right to a higher wage as it studied for a longer time. However, as explained in the previous paragraph, Roemer and Trannoy (2016) argue that the circumstance of these individuals should first be accounted for. As depicted in table 1 and 2, individual D was raised in a rich family with high parental income levels and individual F was raised by parents who received a much lower income. These two individuals belong to different groups as they grew up in different circumstances. A child who grew up with high parental income levels obtains the possibility to study longer as their parents can afford higher intuitions. Therefore, it would be unjust to determine wage based on absolute years studied as individual F never had an equal opportunity. To overcome this problem, Roemer proposed a notion of *relative* effort which is the rank of an individual based on their provided effort given their circumstance. In literature this is known as the RIA – Roemer Identification Assumption (Ramos & Van de Gaer, 2016). The aim of the researcher should be to find a policy that assures individuals D and F receive an equal income as they are both the highest ranked individuals of their respective group as depicted in table 1 and 2 by their percentile score. This means they both provided the highest degree of effort, thus the same *relative* effort, given their circumstance. Note how this is different from strict consequentialism which does not take circumstance into account.

It is important to note that this parental income levels case studied by Roemer and Trannoy (2016) is only one of many circumstances that influence wage. Other possible examples of circumstances include age, gender, a neighbourhood someone grew up in and more. Therefore, a policy that aims for equal wages based solely the circumstance of parental income levels should be considered to be on the lower bound of achieving true equality of opportunity. "Nevertheless, delineating only a few circum-

stances will often suffice to illustrate obvious inequality of opportunity ... and one can then argue that social policy should attempt to mitigate at least that inequality" (Roemer & Trannoy, 2016). The context, nature, scope, and objective of the researcher determines which circumstances and effort variable(s) are relevant to examine. Once circumstances and effort have been determined, the researcher should develop a policy that offset differences due to circumstances but does allow for within-group differences based on different levels of relative effort.

3 METHOD

3.1 *Dataset and task description*

The dataset that this thesis will investigate is known as the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset. This dataset can be retrieved from Propublica Data Store. A total of 7215 instances have been recorded each containing ten different features ('sex', 'age', 'age-cat', 'race', 'priors count', 'juv-fel-count', 'juv-misd-count', 'juv-other-count', 'c-charge-degree', 'c-charge-desc'). A detailed feature description can be found in Appendix A. The target variable is provided in the form of a binary label where '1' represents an individual recommitting a crime and not reoffending is denoted with 0. Y_i is set as the target label. The original data was multi-labelled, but for simplicity the dataset has been bounded by the creators to a binary classification problem.

The task of the algorithm is to predict the target variable y_i (recidivism) as accurately as possible based on the feature vector X_i. Since the target label is known, this will be a supervised learning task. Accuracy, however, is not the only evaluation metric of interest. As the introduction and related work indicated, it is imperative to actively monitor for fairness. The evaluation metric of interest is known as the inequality opportunity ratio (IOR) and will be further explained in section 3.5. This resulted in the following objective: maximizing accuracy of the predicted target variable whilst satisfying the optimal fairness boundaries of the IOR.

Angwin (2018) were the first to discover that the COMPAS dataset was highly biased. The algorithm, that was actually utilized in multiple American states, produced much higher false positive rates for African Americans than for Caucasians. An extensive list of scientific literature yielded similar results for this dataset (Dressel & Farid, 2018; Harrison, Hanson, Jacinto, Ramirez, & Ur, 2020; Lee, Du, & Guerzhoy, 2020). 'Race' will therefore be denoted as the sensitive attribute for this classification

problem. Z_i is a one-dimensional vector containing individual i's race where 1 is the privileged class (Caucasian) and 0 denotes the unprivileged class (African American). Both the feature vector X_i and target label y_i will be provided at the training stage. During the testing phase, y_i will be hidden and only the feature vector X_i can be utilized to form new predictions $\hat{y}_i$. Finally, target label y_i and predicted label $\hat{y}_i$ are compared to evaluate model performance.

3.2 *Data preprocessing*

Within the COMPAS dataset, an individual obtains a 1 if he reoffended within two years and 0 implies the opposite. However, in most binary classification tasks, 1 refers to the positive outcome. Therefore, the labels are substituted at the preprocessing stage. 13.8% of the instances had at least one missing value. Multiple researchers found that it is a better practice to delete instances if the missing data exceeds 10% of the measures (Cismondi et al., 2013; Gorelick, 2006). This is because imputing such a high proportion of missing data would lack natural variation, significantly hurting the validity of the research. After deleting instances with missing values, 6129 entries were left. To all categorical variables ('sex', 'age-cat', 'c-charge-degree') one-hot encoding is applied to create dummy variables. This resulted in 14 possible predictor variables. The dataset is split into a training and test set in a 75:25 ratio respectively using the Scikit-learn module.

3.3 *Analysis - the baseline model*

The prediction task is a straight-forward classification problem in the form of supervised learning. A generalized mixed effects logistic regressor on dataset D will be trained, where D = $\{X_i, y_i\}$. X_i is a vector which contains all the relevant features for individual i. X_i is assumed to be an element to the set of features M, denoted as $X_i \in R^M$.

To select the appropriate features for this particular classification task, an embedded method in the form of Lasso Regression (Least Absolute Shrinkage and Selection Operator) is used. This technique decides which features are relevant to be inputted to the baseline model. Each iteration is evaluated, as it searches for features that contribute the most. Lasso regression performs L1 regularization which penalizes absolute values of the coefficients. If a particular feature is detected as irrelevant, the coefficient is set to zero.

The variable 'race' will be omitted from the input vector X_i as processing of personal data revealing racial origin, producing significant legal effects

on the rights and freedoms of individuals, is prohibited according to article 9 of the GDPR(*General Data Protection Regulation*, 2018, art. 9.). The removal of explicit membership data, such as race, is known in scientific literature as suppression. Therefore, the above mentioned general mixed effects logistic regressor will from now on be referred to as the *suppression baseline model*. All corrective algorithms discussed in section 3.6 will be using the exact same feature input vector. Inputting the same feature vector to all models safeguards an unbiased evaluation of the metrics as all changes occur due differences in models and not due different features.

3.4 *Developing a Roemerian post-processing technique*

To apply the social-economic theory of redistributive justice of Roemer, *effort* and *circumstance* needs to be determined. As explained in paragraph 2.2, effort is "a choice made by the individual, although that choice may be constrained by circumstances, that affects the outcome of interest (Roemer & Trannoy, 2016). In the case of the COMPAS dataset, 'priors-count' can be counted as effort e. The reason that this variable is chosen is because judges regularly consider prior convictions to assess prior effort provided by a criminal to not recidivate. This information is then taken into consideration to decide whether a defendant's sentence should be enhanced (Frase, 2010). The number of prior convictions is therefore considered to be a good indicator, to answer the second research question, and is assumed to be a good exemplifier for the level of effort provided to not recidivate.

A consequentialist would argue that an individual with a higher number of prior convictions should obtain a lower probability to be released. However, as seen in section 2.2, it is important to consider circumstances of individuals as it, for this specific case, may affect the number of prior convictions e. One such a circumstance that might affect effort e is 'race'. An extensive number of scientific papers have found that the number of convictions is related with racial background (Free & Ruesink, 2012; Gross, Possley, & Stephens, 2017; Harmon, 2004). African Americans have significantly more erroneous convictions differentiating from 22% up to 50% more wrongful convictions due police misconduct. Nielsen (2020) ascribes one other possible cause to police discretion in the exercise of how to code an arrest. Another possible cause of these higher prior convictions is redlining - racial discriminatory grading of neighbourhoods – which in turn might lead to more police attention resulting in more arrests for people of colour. To examine whether this higher number of prior count is also present in the COMPAS dataset, the mean difference between both classes will be calculated. Irrespective of the cause, these significantly

higher number of convictions strictly because of race should be accounted for. The following section will describe how the circumstance 'race' is removed from the effort variable 'priors-count'.

The goal is to develop an algorithm that compensates African Americans for the effect of their circumstance on their effort. So, how is it determined whether two individuals from different circumstances provided the same level of *relative* effort? Roemer (1998) proposes to rank individuals within their own group (circumstance). If we take the number of prior convictions for two individuals with different circumstances and they rank in the 99th centile of *their* respective group, they are considered to have provided an equal degree of relative effort. So, if a Caucasian individual is in the top 1% of least prior convictions of his group and an African American is within the top 1% of his respective group, they are considered as having provided an equal level of effort - relative effort - despite the fact that their actual number of prior convictions might differ. This ranking system, offsets the significantly higher number of convictions for African Americans because of their ethnicity.

This algorithm however, will not analyze every percentile separately as the example above, but create a small number of *relative effort* groups instead. Due to the finite nature of datasets it is recommended by Roemer (2013) to choose a circumstance which can take only small range of values. This Roemerian post-processing algorithm will therefore partition the effort variable into three values. How to choose the exact number of values is dependent of the size and nature of the datasets' features, and will be further discussed in the limitation part of chapter 5. All African Americans (unprivileged class) and Caucasians (privileged class) are allocated to either a 'low', 'mid' or 'high' *relative effort* group. Figure 2 illustrates how these relative effort groups are formed.

If an individual from a certain class belongs to the top 33% of lowest prior convictions of his class, this individual is assigned as having provided 'high' effort. Contrary, if an individual from a certain class belongs to the top 33% of most prior convictions of their class, 'low' relative effort is ascribed. This will result in approximately three equally sized groups since 33% of both classes is assigned to one of the three relative effort levels. As explained in section 2.2, equality is achieved when individuals from the same *relative* effort group obtain an equal probability of receiving a positive outcome prediction.

To satisfy the redistributive theory of Roemer, groups with the same level of relative effort should receive an equal opportunity – or probability

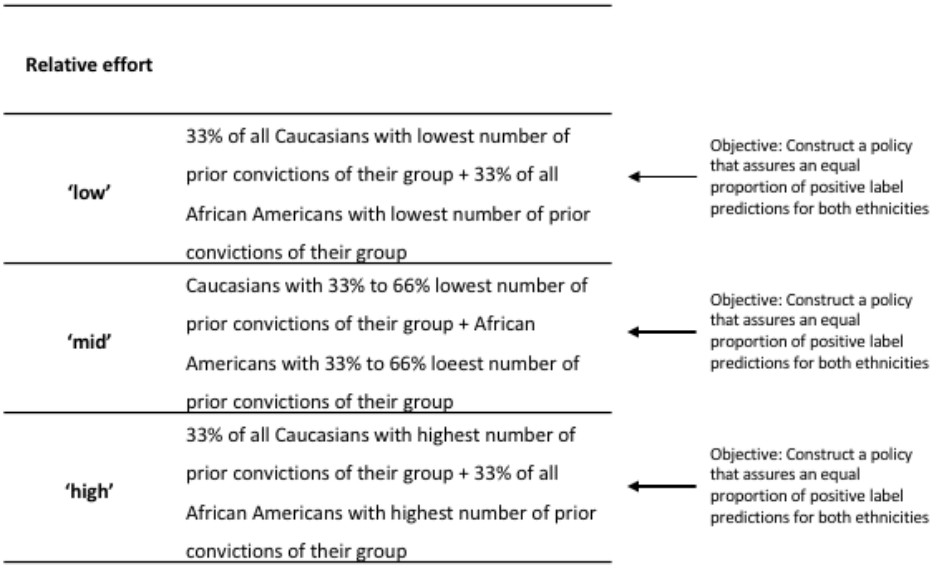

Figure 2: An explanation of the designation of the relative effort groups and the main objective John Roemer's theory of redistributive justice

- of a positive outcome prediction. This implies that the SPD across the privileged class and unprivileged class should be exactly equal. Therefore, the proportions of positive label predictions predicted to the two classes are measured. The predictions generated by the baseline model are evaluated and in case the SPD for any of the 'low', 'mid' or 'high' group deviates from 0, a positive prediction for the privileged class is altered to a negative label prediction. Contrary, a negative prediction for the unprivileged class is turned into a positive outcome prediction. This is an iterative process until the proportions of positive label predictions are exactly equal. However, this post-processing algorithm does not randomly attain an instance from the privileged classes and changes a positive prediction into negative one. It first searches for the closest logarithmic probability just above 0.5. Once it has found the closest logarithmic predictions for the privileged class (e.g. 0.51), it changes the corresponding positive label prediction into a negative one. By searching for close values to 0.5, this algorithm only alters predictions outcomes that initially were difficult to predict. Turning a value of 0.51 to a 0-label prediction is less likely to affect accuracy than changing a random value (e.g. 0.99) to 0. The same logic applies for altering a negative prediction label into a positive label prediction.

3.5 *Evaluation metrics*

To measure to what extent this method achieves equality opportunity, and thereby answering RQ3, various metrics could be adopted. The general entropy measures can be rewritten such that the cumulative income distribution functions represent the sum of inequalities across types. Additionally, several scientific papers defined ratios to measure inequality due circumstances (Checchi & Peragine, 2010; Ferreira, Gignoux, & Aran, 2011; Roemer, 2013). However, this thesis will follow the metric defined by Roemer Trannoy (2016) which is most practical in case circumstance is already defined:

$$\eta = y^d / \overline{y} \tag{7}$$

$\overline{y}$ is the average income across all types whereas y_d represents the income of the most disadvantageous type. The optimal objective is to achieve an η of 1, but equality of opportunity is achieved if it suffices the fairness boundaries of 0.9 and 1.1.

However, in this case the main objective is not to examine income, but recidivism instead. To evaluate performance of the Roemerian post-processing algorithm, the inequality opportunity ratio (IOR) is slightly adjusted to enable the measurement of the binary target label recidivism. The IOR divides the predicted positive labels for the unprivileged class by the average of predictive positive label across all classes. This resulted in the following fairness metric:

$$IOR = \frac{P(\hat{y} = 1|z = 0)}{P(\hat{y} = 1)} \tag{8}$$

To present an unbiased view of the algorithm created, other fairness definitions (see section 2.1) are considered in the results as well. Statistical parity difference, disparate impact, equal opportunity difference, equalized odds, FPR and FNR respectively will be evaluated.

3.6 *Model performance comparison*

To answer the last research question and examine how well the Roemerian post-processing technique performs against the bias mitigation algorithms of chapter 2.1, they first have to be recreated. These techniques will be implemented using the AIF360 module in Python (Bellamy et al., 2018). The following corrective algorithms are considered: Reweighing (Kamiran & Calders, 2012), Learning fair representations (Zemel, Wu, Swersky, Pitassi, & Dwork, 2013), Adversarial debiasing (Zhang et al., 2018) and Equal odds postprocessing (Hardt et al., 2016). All models will obtain the exact

same feature input vector as the baseline model. The AIF360 module from Bellamy et al. (2018) was constructed in such a manner that these models only need information about the sensitive attribute (race) and access to the training and test data to generate predictions. Adversarial debiasing, however, is the only exception as it requires three hyperparameters to be tuned. GridSearchCV was imported from the Scikit-Learn library. The hyperparameters are optimized using 5-fold cross validation on the training data X_{train}. The accuracy score on a dedicated evaluation set (a randomly hidden proportion of X_{train} that was not seen during the model selection step) will be calculated. The work of Goodfellow, Bengio, and Courville (2016) is followed in which they recommend to evaluate a range of grid search values. They advocate a batch size between 32 and 512 to compute the approximation to the gradient. Potential learning rates is suggested to range between 0.1 and 0.1^{-6} and the number of epochs advocated is somewhere between 10 and 150.

4 RESULTS

4.1 *Lasso analysis for the baseline model*

To develop the baseline model, Lasso regression was performed to find relevant input variables. A total of ten different features were selected. However, the variable 'race' will be omitted as input variable. This is a requirement set by the GDPR in article 9 (*General Data Protection Regulation*, 2018, art. 9.). This is especially the case for automated processing involving life-altering decisions such as sentencing guidelines based on the likelihood of recidivism occurring (Feldman et al., 2015). Omitting the sensitive variable of interest is known as suppression in scientific literature. Therefore, the *suppression baseline model* resulted in a general mixed effects logistic regressor with the following features as input: 'sex-Female', 'sex-Male', 'age-cat-Greater-than-45', 'age-cat-Less-than-25', 'priors-count', 'juv-fel-count', 'juv-misd-count', 'juv-other-count', 'c-charge-degree', 'c-charge-desc-F'. The associated coefficient values can be found in Appendix B.

4.2 *Roemerian post-processing analysis*

In this section the suggested methodology of section 3.4 is executed and compared to the baseline model. The first step consisted of separating instances based on circumstance c. Circumstance c was set to be the variable 'race', resulting in an African American class and a Caucasian class. Both classes were divided to either a high, mid or low relative effort

group separated by the count of prior convictions *e*. Table 3 provides some interesting statistics per class.

Remember that multiple scientific research papers discovered a higher number of prior convictions for African Americans due to either more erroneous convictions, redlining or police discretion (Free & Ruesink, 2012; Gross et al., 2017; Harmon, 2004; Nielsen, 2020). Table 1 illustrates that these results are in line with prior literature. African Americans have a significantly higher average of 4.45 prior convictions compared to an average of 2.59 for Caucasians. The respective 33-th and 66-th percentile differ significantly for both classes as well.

Number of prior convictions (e)	African Americans (c)	Caucasians (c)
Average	4.45	2.59
33% 'low' relative effort	$e \geq 5$	$e \geq 2$
33% 'mid' relative effort	$1 < e < 5$	$0 < e < 2$
33% 'high' relative effort	$e \leq 1$	$e \leq 0$

Table 3: Assignment of relative effort based on prior convictions per class (N=6129)

In order to answer RQ2, the following question should be answered: "What variable(s) exemplifies the degree of effort provided by an individual to receive a positive outcome?". The number of prior convictions is regularly used by judges to decide whether a defendant's sentence should be enhanced, and rightfully so. The Lasso regression model selected it as relevant feature for predicting recidivism as well. The results also showed that the number of prior convictions is seriously affected by the circumstance 'race', indicated by the notable mean group difference. The definition of effort founded in 2.2 is thus satisfied as the number of prior convictions can indeed be observed as "a choice made by the individual, although that choice may be constrained by circumstances, that affects the outcome of interest".

In order to answer research question 3, to what extent the Roemerian post-processing method can achieve equal opportunity, the suggested methodology of 3.4 first has to be developed. Three different relative effort groups are created and all individuals are assigned to one of these groups. When a particular individual is assigned to a group can be found in table 3. For example, an African American who has committed 5+ foregoing crimes is assigned as having provided a *low* level of relative effort whereas a Caucasian is already assigned to *low* levels of relative effort when 2+ preliminary crimes were committed. After all relative effort groups have been formed, the algorithm alternates classification labels until the SPD across both classes are equal (see section 3.4).

The significantly higher average of prior convictions for African Americans illustrate why it is of vital importance to assign groups based on *relative* effort and not on absolute levels of effort. If race was not accounted for, significantly more African Americans would have ended up in the *low* effort group. Because effort is allowed to be a variable of influence on the outcome an individual receives, it would have yielded worse outcomes for relatively more African Americans.

To understand the impact of the post-processing algorithm on the classification task, the test labels were evaluated. Table 4 shows the predicted positive labels for the ground labels, predicted labels from the baseline model and the predicted label from the Roemerian post-processing.

(Predicted) Positive Labels	African American	Caucasian
Ground labels (Y_{test})	447/911=49.1%	371/622=59.6%
Suppression baseline	493/911=54.1%	472/622=75.9%
Roemerian post-processing	483/911=53.0%	290/622=46.6%

Table 4: Percentage of positive labels predicted over the test (N=1533)

(Predicted) Positive Labels	African American
Ground labels (Y_{test})	49.1/59.6=0.824
Suppression baseline	54.1/75.9=0.713
Roemerian post-processing	53.0/46.6=1.137

Table 5: Disparate impact measured in % predicted positive labels per class (N=1533)

The suppression baseline model had an increase of 46 positive African American label predictions and 101 positive Caucasian label predictions compared to the ground truth. This significantly larger increase for Caucasians resulted in a disparate impact ratio of 0.713 dropping below the legal support levels of 0.8 (*EEOC compliance manual.*, 1978). Disparate impact is shown in table 5. The suppression model actually amplified the discrimination in the dataset since the disparate impact of the ground truth labels was higher at 0.824. Results are in line with prior research which showcase suppression can still return discriminating outputs (Besse, del Barrio, Gordaliza, Loubes, & Risser, 2020; Corbett-Davies et al., 2017).

The corrective Roemerian post-processing algorithm has increased 36 positive African American labels but had a decrease of 81 positive label predictions for the Caucasian class. Disparate impact therefore increased to 1.137 satisfying the optimal prevalence difference boundaries of 0.8 and 1.2 as determined in chapter 2.1. A disparate impact exceeding 1 designates that the proportion of positive label predictions of African Americans exceeded the percentage of positive label predictions predicted for Caucasians.

In order to provide a concrete answer to research question 3, to what extent the Roemerian model achieves algorithmic fairness, the IOR is evaluated. Complementary, multiple independent fairness criteria are analyzed as well. Figure 3 compares the suppression baseline to the post-processing methodology. The independent criteria will be Inequality Opportunity ratio, Statistical Parity difference, Disparate Impact, Equal Opportunity difference, Equalized odds, False Positive rate and False Negative rate.
The suppression baseline method fails to satisfy all boundaries set in

Figure 3: : Performance of the suppression and corrective post-processing algorithm on the fairness metrics defined in chapter 2

chapter 2.1. This shows why it is necessary go beyond GDPR guidelines and to adopt fairness enhancement techniques. Contrary, every single fairness metric moves closer to the objective line after the Roemerian post-processing. Six of the seven fairness metrics move within the objective boundaries. The inequality opportunity ratio significantly improves from 0.86 to 1.051 after post-processing. This advancement provides African Americans with 5.1% more positive label predictions compared to the average across all individuals in the test set. Research question 3 can therefore be answered positively as the model successfully applied the theory of Roemer to obtain equality of opportunity, measured by the IOR, as it falls within the objective fairness boundaries of 0.9 and 1.1.

4.3 *Performance of all fairness classifiers*

To answer the last research question, it should be evaluated how well this Roemerian post-processing performs against the popular fairness enhancement algorithms explained in section 2.1. Fairness and accuracy is often displayed as a zero-sum game where an increase in fairness would

result in a lower accuracy. Since all fairness metrics improved after the Roemerian post-processing algorithm, it will be interesting measure what effect this corrective post-processing has had on accuracy. Additionally, the following bias mitigation algorithms will be evaluated: Reweighing (Kamiran & Calders, 2012), Learning fair representations (Zemel et al., 2013), Adversarial debiasing (Zhang et al., 2018) and Equal odds post-processing (Hardt et al., 2016). All these models obtained the same input vector as the baseline model to assure an unbiased comparison. The AIF360 module created by Bellamy et al. (2018) enabled a straightforward implementation of these fairness methods as they exclusively required information about the sensitive attribute 'race' and access to the training and test set. Only the GAN, in the form of Adversarial debiasing, required three hyperparameters to be tuned. The accuracy score of all combinations of the grid search values is displayed in Appendix C. The optimal learning rate was found at 0.01, the optimal number of epochs were determined at 100 and the batch size was set at 128. An overview of the optimized hyperparameters are illustrated below in table 6.

Model	Feature description	Grid search value	Optimal value
Adversarial debiasing	Learning rate	0.001, 0.01, 0.1	0.01
	Number of epochs	25, 50, 100	100
	Batch Size	32, 64, 128	128

Table 6: Overview of tuned hyperparameters and the optimal value adopted for Adversarial debiasing

After all these bias mitigation techniques have been created, the results can be examined. Table 6 illustrates whether the obtained results fall within the fairness boundaries set in chapter 2.1 and 3.6. Green values indicate that they are within their boundary whereas red values fail to meet the objective boundaries. To determine whether the achieved accuracy score of the classifier are relevant, they are compared to a Zero rate classifier. This classifier always predicts the most frequent class present in the test set. Accuracy scores above the Zero Rate classifier's accuracy score of 53.4% for this COMPAS dataset, are therefore coloured green in table 7. A machine learning model should always be above the so-called ZeroR accuracy score to be deemed useful for predictions (Brownlee, 2019).

The original accuracy of the baseline model was 67.2%. The Roemerian corrective post-processing dropped accuracy to 63.0%. This resulted in a drop of 4.2%. However, this drop in accuracy is arguably a good trade-off to make, as it results in an improvement on all fairness metrics. This

	Accuracy	Inequality opportunity ratio	Statistical parity difference	Disparate impact	Equality of opportunity	Equalized odds	False positive rate	False negative rate
Objective	**1**	**1**	**0**	**1**	**0**	**0**	**1**	**1**
Suppression	0.672	0.860	-0.218	0.713	-0.137	-0.191	1.665	0.511
Roemerian postprocessing	0.630	1.051	0.064	1.137	0.065	0.096	0.691	1.189
Reweighing	0.660	0.998	-0.004	0.994	0.063	0.025	1.029	1.334
Learning fair representations	0.382	0.997	-0.004	0.994	-0.036	-0.029	1.034	0.938
Adversarial debiasing	0.665	0.975	-0.037	0.941	-0.005	-0.004	1.004	0.978
Eqodds postprocessing	0.620	0.986	-0.025	0.966	-0.004	-0.002	1.001	0.981

Table 7: Objective and performance scores of all models on accuracy and fairness metrics

assures that the model does not stimulate real-life discriminatory decision-making. A noteworthy observation is that the post-processing algorithm actually has the most excessive increase on the inequality opportunity ratio, statistical parity difference, disparate impact, equality of opportunity difference and equalized odds. It has actually surpassed the objective values of these metrics. The African American group receives a slightly higher proportion of positive label predictions as compared to the Caucasian group. These results depict that the African American group are now in a privileged condition, flipping the privileged and unprivileged class. However, these fairness metrics are still well within the objective boundaries as earlier discussed and illustrated in figure 1.

Finally, the scores of table 7 are summarized into figure 4 visually displaying the differences between models. The highest accuracy achieved as compared to the suppression baseline is Adversarial debiasing which only showed a slight decrease of 0.7% with an accuracy score of 0.665. This neural networking technique also managed to satisfy all objective fairness boundaries making it the most sufficient fairness enhancement technique. Numerous scientific papers have produced similar results demonstrating its' incredible prediction power whilst optimizing multiple fairness metrics (Wadsworth, Vera, & Piech, 2018; Zhang et al., 2018). This thesis adds to existing scientific literature as it finds complementary results with respect to the COMPAS dataset. Equalized odds and learning fair representation also satisfied all fairness metrics but had a significantly larger drop in accuracy with the LFR algorithm dropping far below the accuracy of the majority classifier. Reweighing only failed to satisfy the boundary for

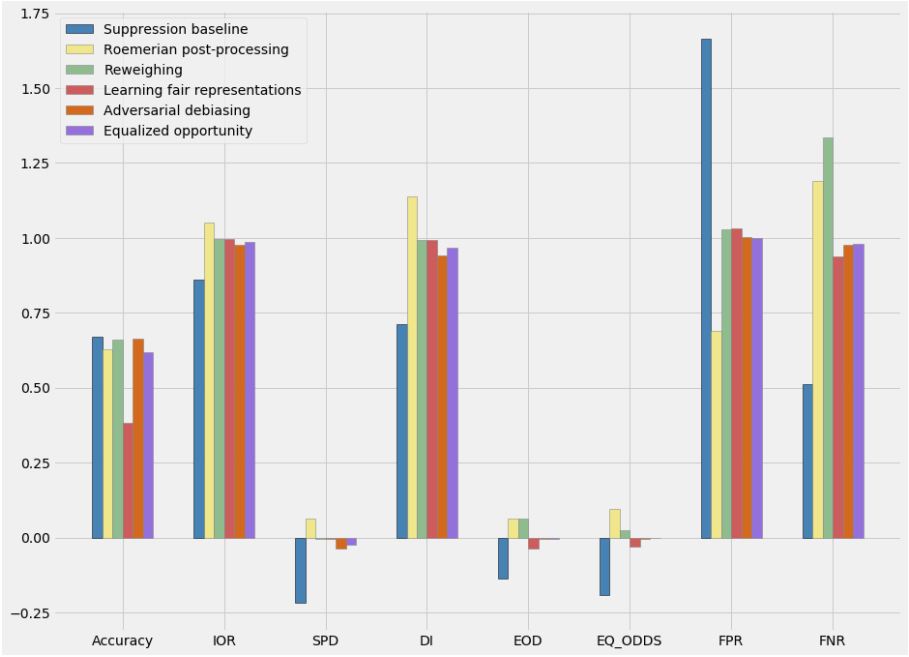

Figure 4: : Performance of fairness-aware algorithms on accuracy and fairness metrics defined in chapter 2.1

the FNR but did produce accurate predictions, coming in second with an accuracy score of 0.660. Figure 4 shows that fairness metrics can drastically be improved without a significant drop in accuracy.

Conclusively, and in response to research question 4, how does the Roemerian model perform compared to other bias mitigation algorithms previously developed in the algorithmic fairness domain? It outperforms all fairness metrics compared to the suppression baseline. It fails to meet the FPR fairness boundary but meets the other six objectives. Only Equalized odds post-processing and Adversarial debiasing succeeded to satisfy all metrics. However, the Roemerian model does outperform equalized odds post-processing in terms of accuracy. The outstanding method is Adversarial debiasing as it obtains the highest accuracy of all bias mitigation algorithms whilst satisfying all fairness boundaries. Despite their staggering performance, the opacity of neural networks is a real concern. The decision as to why a certain label is predicted is non-transparent and therefore have lost some of the substantive connection to society's conception of justice. The Roemerian algorithm, however, maintains a strong connection by taking steps and decisions in line with John Roemer's social-economic theory of redistributive justice.

5 DISCUSSION

5.1 *Theoretical and practical implications*

The goal of this study was to examine to what extend a social economic theory of redistributive justice developed by Roemer could be applied to the COMPAS dataset to increase algorithmic fairness. Results show that the theory can be successfully applied as it satisfies six of the seven fairness metrics. It also illustrates that incorporating bias mitigation algorithms does not necessarily have to lead to a dramatic decrease in accuracy with the newly created Roemerian post-processing dropping 4.2% and Adversarial debiasing neural network dropping just 0.7%. To the best of my knowledge, it is one of the first works that explores a social-economic theory to predict a non-economic outcome of interest, namely recidivism. Both theory and results also illustrated that suppression, the exclusion of sensitive attributes is not merely enough to prevent discriminatory outputs despite this being the most widely adopted technique by businesses (Groen, 2018). This result is in line with a multitude of scientific research papers (Barocas & Selbst, 2016; Hajian et al., 2016). It stresses the importance for businesses to pursue their own qualitative research of social-economic theories and the necessity to develop new fairness enhancing algorithm. In doing so, a strong connection between the developed method and normative conception of justice in society remains.

5.2 *Limitations*

Surprisingly, the overall disparate impact turned out to be 1.137 moderately over the objective goal of 1. This deviation is due to different group proportions for the relative effort classes. Contrary, to Roemers' framework which split different groups based the continuous variable annual income levels, this thesis split groups based on the number of prior convictions, a discrete variable. With a continuous variable, it is less complex to choose cut-off points that create roughly equal group. However, due to the finite and small range of values for the number of prior convictions, the number of groups should be carefully chosen. This thesis encountered this problem as it started out by creating 5 different relative effort groups. However, as it turned out, the 20^{th} and 40^{th} percentile for the Caucasian group was both at a prior convictions count of 0. Because these Caucasian individuals have the exact same of relative effort provided it would be unfair to arbitrarily split Caucasian into one of the two groups. This means the algorithm would have ended up with one empty 40^{th} percentile group for the Caucasian which makes it impractical to operate on. Additionally, relatively more

Caucasian would have been assigned to the 'high' relative effort group and therefore disproportionally more Caucasians would obtain a positive label prediction resulting in worse fairness metrics. Therefore, the Roemerian algorithm, with respect to the COMPAS dataset, chose to partition three relative effort groups 'low', 'mid' and 'high'. It is of crucial importance to carefully consider the number of percentile splits when working with a discrete effort variable and this remains an interesting topic for future work.

Furthermore, Roemer his redistributive justice theory as well as the developed algorithm assumes no conditional dependency between the sensitive attributes and the outcome obtained. This assumes no difference in underlying distribution of effort between the African American and Caucasian class. Whether this is the right assumption differs for each context and remains a part of philosophical debate. However, in certain situations it might be imperative evaluate and assume possible correlations between the outcome and the protected attributes. In medical treatments for example, race can be a crucial determinant in the effectiveness of a drug. If a medicine would work significantly better for a certain protected class, this drug should be distributed disproportionally to this class. It would be unethical to optimize for statistical parity to provide the same proportions of treatments to both classes at the expense of human lives. In such cases, disregarding a possible connection between sensitive attributes and outcomes would lead to a humongous unethical result.

5.3 *Future work*

Future work can focus on the problem described above regarding the division of relative effort groups when working with discrete variables. How can we choose the optimal number of groups with respect to the circumstance and effort chosen. An interesting domain to explore is to perhaps develop a rule of thumb or finding potential proxies or other techniques to deal with discrete effort variables. Furthermore, this thesis considered only one circumstance that affects the effort variable prior convictions. The same logic applies to the effort variable whereas, depending on the context, perhaps multiple variables can be combined to measure the degree of effort. Future work is encouraged to extend this method by introducing measurement of circumstances and effort in a multi-dimensional context. Lastly, a multitude of contrasting egalitarian theories exist in scientific literature. Noteworthy theories to examine have been developed by Robert Nozick, Kenneth Arrow John Harsanyi (Arrow, 1973; Harsanyi, 1975; Nozick, 1973). It would be valuable if these social-economic theories are explored as well in order to create algorithms that achieve algorithmic fair-

ness whilst maintaining a robust connection to concepts of justice present in society.

6 CONCLUSION

This thesis proposes a new corrective post-processing algorithm incorporating a social-economic fairness framework of 1998 created by John Roemer (Roemer, 1998). Results show that the developed algorithm improved all fairness metrics compared to the suppression baseline, and, satisfies six out of the seven predefined fairness metrics. To develop and evaluate this post-processing technique, four auxiliary research questioned were answered. First, prevailing fairness metrics and widely adopted techniques were examined. Secondly, a potential effort variable was designated. It demonstrated how the number of prior convictions, in line with prior literature, is constrained by the circumstance 'race'. To remove the influence from circumstance on effort, a post-processing method was developed which alters predictions until the SPD is equalized across the African American and Caucasian *relative* effort groups. This technique was then compared to the bias mitigation techniques previously mentioned and tested on a range of different fairness metrics. It illustrated that these techniques does not necessarily have to lead to a dramatic decrease in accuracy. However, due to the complexity and non-transparency of these techniques, they lost a substantive connection to theories and fairness as it is conceived of in society. By developing an algorithm in line with the reasoning of a social-economic theory of redistributive justice, a robust connection remains. Results therefore show that John Roemer's theory of redistributive justice can successfully be applied as algorithmic fairness.

REFERENCES

Angwin, L. J. M. S. . K. L., J. (2018). Machine bias. *Investigative Journal of Journalism*.

Arrow, K. J. (1973). *Some ordinalist-utilitarian notes on rawls's theory of justice.* JSTOR.

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

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

Besse, P., del Barrio, E., Gordaliza, P., Loubes, J.-M., & Risser, L. (2020). A survey of bias in machine learning through the prism of statistical parity for the adult data set. *arXiv preprint arXiv:2003.14263*.

Brownlee, J. (2019). Machine learning mastery with weka. *Ebook. Edition*, *1*, 4.

Calders, T., & Verwer, S. (2010). Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, *21*(2), 277–292.

Checchi, D., & Peragine, V. (2010). Inequality of opportunity in italy. *The Journal of Economic Inequality*, *8*(4), 429–450.

Cismondi, F., Fialho, A. S., Vieira, S. M., Reti, S. R., Sousa, J. M., & Finkelstein, S. N. (2013). Missing data in medical databases: Impute, delete or classify? *Artificial intelligence in medicine*, *58*(1), 63–72.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 797–806).

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, *4*(1), eaao5580.

Eeoc compliance manual. (1978). Washington, D.C.]: U.S. Equal Employment Opportunity Commission.

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

Ferreira, F. H., Gignoux, J., & Aran, M. (2011). Measuring inequality of opportunity with imperfect data: the case of turkey. *The Journal of Economic Inequality*, *9*(4), 651–680.

Frase, R. S. (2010). Prior-conviction sentencing enhancements: Rationales and limits based on retributive and utilitarian proportionality prin-

ciples and social equality goals. *Previous convictions at sentencing: Theoretical and applied perspectives*, 117–136.

Free, M. D., & Ruesink, M. (2012). *Race and justice: Wrongful convictions of african american men*. Lynne Rienner Publishers Boulder, CO.

Garg, P., Villasenor, J., & Foggo, V. (2020). Fairness metrics: A comparative analysis. In *2020 ieee international conference on big data (big data)* (pp. 3662–3666).

General data protection regulation. (2018). Retrieved 2021-02-03, from https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf

Ghassami, A., Khodadadian, S., & Kiyavash, N. (2018). Fairness in supervised learning: An information theoretic approach. In *2018 ieee international symposium on information theory (isit)* (pp. 176–180).

Giannotti, F., & Pedreschi, D. (2008). *Mobility, data mining and privacy: Geographic knowledge discovery*. Springer Science & Business Media.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Gorelick, M. H. (2006). Bias arising from missing data in predictive models. *Journal of clinical epidemiology*, *59*(10), 1115–1123.

Groen, B. A. (2018). A survey study into participation in goal setting, fairness, and goal commitment: Effects of including multiple types of fairness. *Journal of Management Accounting Research*, *30*(2), 207–240.

Gross, S. R., Possley, M., & Stephens, K. (2017). Race and wrongful convictions in the united states.

Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 2125–2126).

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, *29*, 3315–3323.

Harmon, T. R. (2004). Race for your life: An analysis of the role of race in erroneous capital convictions. *Criminal Justice Review*, *29*(1), 76–96.

Harrison, G., Hanson, J., Jacinto, C., Ramirez, J., & Ur, B. (2020). An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 392–402).

Harsanyi, J. C. (1975). Can the maximin principle serve as a basis for morality? a critique of john rawls's theory. *American political science review*, *69*(2), 594–606.

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

Jiang, H., & Nachum, O. (2020). Identifying and correcting label bias in machine learning. In *International conference on artificial intelligence and statistics* (pp. 702–712).

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

Lee, C. S., Du, J., & Guerzhoy, M. (2020). Auditing the compas recidivism risk assessment tool: Predictive modelling and algorithmic fairness in cs1. In *Proceedings of the 2020 acm conference on innovation and technology in computer science education* (pp. 535–536).

Najibi, A. (2020). *Aracial iscrimination in face detection technology.* Retrieved from https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

Nielsen, A. (2020). *Practical fairness.* O'Reilly Media.

Nozick, R. (1973). Distributive justice. *Philosophy & Public Affairs*, 45–126.

Ramos, X., & Van de Gaer, D. (2016). Approaches to inequality of opportunity: Principles, measures and evidence. *Journal of Economic Surveys*, *30*(5), 855–883.

Rawls, J. (1958). Justice as fairness. *The philosophical review*, *67*(2), 164–194.

Roemer, J. E. (1998). *Theories of distributive justice*. Harvard University Press.

Roemer, J. E. (2013). Economic development as opportunity equalization. *The World Bank Economic Review*, *28*(2), 189–209.

Roemer, J. E., & Trannoy, A. (2016). Equality of opportunity: Theory and measurement. *Journal of Economic Literature*, *54*(4), 1288–1332.

Wadsworth, C., Vera, F., & Piech, C. (2018). Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199*.

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

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *International conference on machine learning* (pp. 325–333).

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 aaai/acm conference on ai, ethics, and society* (pp. 335–340).

Zhong, Z. (2018). *A tutorial on fairness in machine learning.* Retrieved from https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

APPENDIX A

Feature abbreviation	Feature description
X_{sex}	Sex of the data subject
X_{age}	Age in the number of years of data subject
$X_{age\text{-}cat}$	Age category of data subject (<25,25-45,>45)
X_{race}	Race of the data subject
$X_{priors\text{-}count}$	Number of prior convictions of data subject
$X_{juv\text{-}fel\text{-}count}$	Binary variable indicating whether a data subject committed a felony before the age of 18
$X_{juv\_misd\_count}$	Binary variable indicating whether a data subject committed a misdemeanour before the age of 18
$X_{juv\text{-}other\text{-}count}$	Binary variable indicating whether a data subject committed an infraction or other type of crime before the age of 18
$X_{c\text{-}charge\text{-}degree}$	Categorial variable that indicates whether it is a felony, misdemeanour or other type of crime
$X_{c\text{-}charge\_desc}$	Detailed description of the type of crime

Table A. 1: Descriptions of all features available in the COMPAS dataset

APPENDIX B

Variables	Coefficients	Alpha
Juv-fel-count	-0.023739	0.000682
Juv-misd-count	-0.000447	
Juv-other-count	-0.035139	
Priors-count	-0.028711	
Sex-female	0.064067	
Race	0.019133	
Sex-male	-0.002885	
Age-cat-25-45	0	
Age-cat-greater-than-45	0.147384	
Age-cat-less-than-25	-0.157862	
C-charge-degree-f	-0.035878	
C-charge-degree-m	0	

Table B. 1: Feature importance of all features available in the COMPAS dataset using Lasso Regression model implemented from the SKlearn Module

APPENDIX C

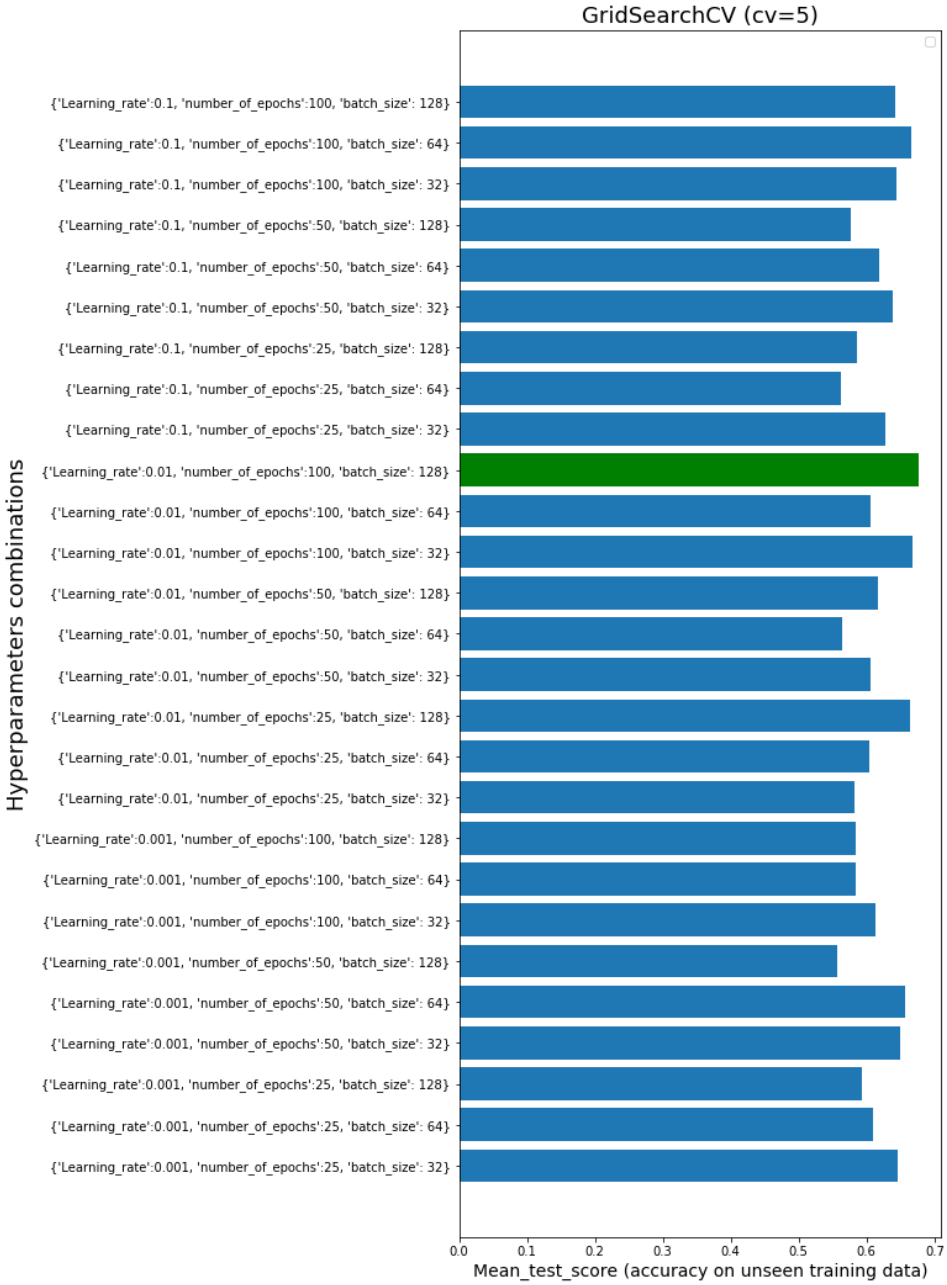

Figure C. 1: Bar plot of the accuracy score during the grid search for the Adversarial debiasing method