Fairness in Machine Learning: Methods for Correcting Black-Box Algorithms

Fairness in Machine Learning

Methods for Correcting Black-Box Algorithms

A thesis presented

by

Amil Merchant

to

Applied Mathematics

in partial fulfillment of the honors requirement

for the degree of

Bachelor of Arts

Harvard College

Cambridge, Massachusetts

March 29, 2019

Table of Contents

Abstract

Machine learning is being used more frequently across a wide range of social domains. These algorithms are already trusted to make impactful decisions on topics including loan grades, personalized medicine, hiring, and policing. Unfortunately, many of these models have recently been criticized for discrimination against individuals of different races or sexes. This is particularly problematic from a legal perspective and has led to challenges over the use of these algorithms. In this thesis, we consider what would be needed to make a machine learning model fair according to the law. Special emphasis is placed on the COMPAS algorithm, a black-box machine learning model used for criminal recidivism prediction that has recently been shown to have a discriminatory impact for defendants of different races. We test two algorithmic methods in adversarial examples and adversarial networks that show significant progress in meeting the proposed legal requirements of fairness.

Acknowledgements

First, I am indebted to my advisor Michael Brenner for first introducing me to this problem and his support throughout. His insights into formulating research problems, the field of fairness, and how to frame this thesis have been invaluable.

Also, this work would not have been possible without all of our research collaborators. Special thanks go to Suproteem Sarkar, Drew Wegner, Michael Haley, Richard Millett, Matthew Lin, and Jeannie Suk Gersen for all of their help in understanding the legality of machine learning algorithms. I was amazed by their dedication to exploring this problem, and I am so grateful for all our early-morning discussions that really helped bridge the legal and technical perspectives on fairness.

I would also like to thank Sean Eddy for taking the time to read this thesis and David Gibson for his helpful advice.

Part I

Introduction to Fairness

Chapter 1

Introduction

1.1 Motivation

Machine learning is being increasingly adopted within a wide range of real-world applications. Learning from historical data, these algorithms showcase a remarkable ability to uncover statistical relationships overlooked by humans and generalize trends between observed characteristics and social outcomes [14]. High prediction accuracies have been noted for a variety of tasks, including credit scoring [26, 27] and personalized medicine [36]. However, this form of evidence-based learning is far from perfect.

Fairness is one of the most prominent concerns, and many models have recently come under criticism for alleged discrimination. Since training data often includes historical prejudices, algorithms repeat these biases within predictions. For example, Amazon recently faced criticism for a resume-screening algorithm that was biased against female applicants in a traditionally male-dominated industry [13]. Recent case studies have also examined how algorithms may lead to differential pricing or even prevent certain individuals from obtaining loans [34]. Clearly, machine learning models that incorporate and relay such biases can have dangerous social implications.

Even worse, this discriminatory impact is likely to worsen over time due to self-fulfilling effects and feedback loops. Consider the problem of predictive policing. A recent study of the PredPol algorithm used to allocate law enforcement to certain neighborhoods found that the model not only targeted African-American communities at the beginning but also increasingly targeted these areas over time, even though the rates of crime were the same [32, 10]. As machine learning algorithms are increasingly used for such high-impact decision making, it is necessary to consider how these models can formalize prejudice and what can be done to correct predictions.

1.2 Criminal Recidivism Prediction: COMPAS

A particularly concerning use case of machine learning is in criminal recidivism prediction. In the United States, algorithmic predictions are being increasingly incorporated in all phases of the criminal justice system, including bail, sentencing, and parole decisions [9]. The risk component of these algorithms attempts to predict the probability of a repeat offense, and high scores are intended to inform judges that defendants should be denied bail or deserve harsher sentencing. The potential life-altering impact of these predictions compels further study to determine if these models are fair[1] [16, 10].

In this thesis, we focus on a popular model known as the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm which is used in at least 5 states as a factor in bail and sentencing decisions[2] [1]. The algorithm is purported to create a psychological model of the individual to predict the risk of recidivism [5]. To do so, the machine learning model relies on a 137-part questionnaire, examples of which are seen in figure 1.1. While some of the questions on current charges or criminal history are obviously probative, others on personality traits, family/friends, and sociocultural factors are very concerning. The answers could easily incorporate historical prejudices and demographic differences into the COMPAS model and lead to discrimination based on race or sex.

4. Based on the screener's observations, is this person a suspected or admitted gang member?

65. Is there much crime in your neighborhood?

98. Do you feel discouraged at times?

127. A hungry person has a right to steal. [Agree or Disagree]

Figure 1.1: Example Questions from COMPAS [3]

Until 2016, this model was utilized without significant examination from the greater research community. The model was generally treated as a 'black-box.' States did not pay for the algorithm itself but rather bought a subscription to the evaluation service from the parent company Northpointe (which has since been sold to Equivant)

[1] A recent study by Dobbie et al. notes also confirms the self-fulling nature of recidivism prediction. If a defendant is denied bail, there is a significantly higher probability of conviction. This is despite the fact that detention has no net effect on crime. This feedback loop could confirm racial biases within predictions and further supports why models need to be fair in the first place.

[2] Including Florida, Michigan, New Mexico, Wisconsin, and Wyoming

[4]. Therefore, the model parameters were and still are not known, fundamentally limiting how the model can be validated [37, 23].

Finally, in a groundbreaking study, the investigative journalism group ProPublica utilized Freedom of Information Act requests from Broward County in Florida to obtain details about various cases and the outputted COMPAS risk scores (but not the model or any survey answers as these remained proprietary and confidential). Their analysis revealed a generic trend of discrimination based on race across the entire ProPublica dataset. Black defendants were significantly more likely to be given a higher score, as seen in figure 1.2, highlighting the discriminatory problem of COMPAS[3].

Northpointe and the states using these algorithms have denied this bias, and instead, they promote how the model predictions reflects the true probability of recidivism Since ProPublica's analysis, an ongoing debate has surrounded COMPAS. It is now one of the fundamental problems within the field of fairness in machine learning, and the legal and computer science literature continues to struggle to define if this algorithm is fair and, if not, what corrections should be made [37, 11].

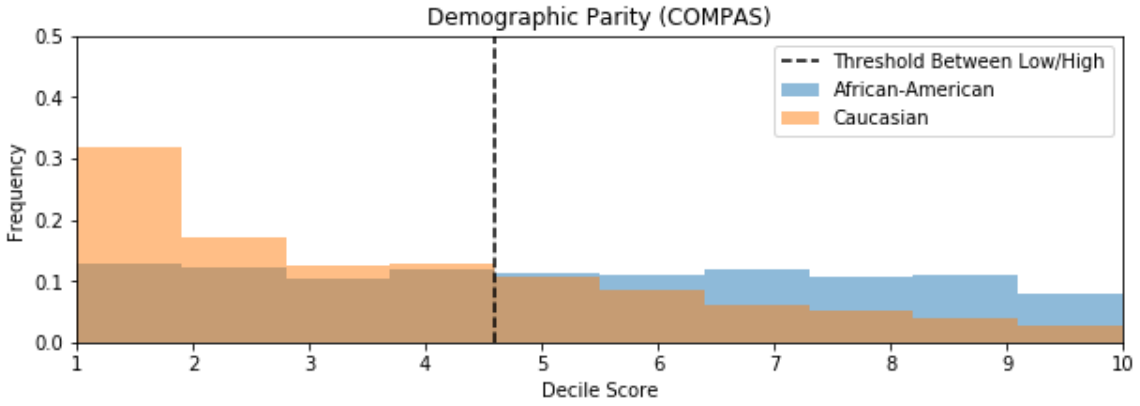


Figure 1.2: Score Distributions of COMPAS

1.3 Outline

This thesis is a technical extension of on-going work with researchers at the Harvard Law School to explore the legality of COMPAS and other machine learning models within the justice system. Current computer science literature has proposed various

[3]For example, the authors compared two DUI arrests. An African-American defendant with no history of DUIs was given a medium risk, whereas a Caucasian man was given the lowest possible score even though he had a history of at least 3 previous DUIs. In many such cases, the predictions were inaccurate and the defendants with lower scores ended up being the ones to recidivate.

mathematical definitions of fairness and methods to satisfy the constraints; however, little work has been done to see if the same contributions would uphold from a legal perspective. The purpose of this thesis is to bridge the gap between the technical and legal arguments. Working with the law students, we asked what would make a model fair, and the methods presented in this thesis examine if algorithmic methods that enforce fairness can meet the legal definitions.

To even start this examination, we need a machine learning model to test. Faced with the unavailability of COMPAS parameters[4], chapter 2 describes the use of the ProPublica dataset to create a reconstruction of Northpointe's algorithm based on the available features. A simple linear model is helpful in confirming the discrimination described by ProPublica but does not capture the intricate correlations between feature variables. Instead, we then turn to traditional black-box reconstruction techniques. Specifically, a neural network is trained using the available data from ProPublica to mimic risk scores. This model serves as a proxy for COMPAS, and it is the basis for the algorithmic methods of fairness in this thesis.

Next, chapter 3 introduces the mathematical formalization of three common definitions of fairness within the computer science literature and metrics used to measure them. Current legal precedent supports COMPAS and creating models that are calibrated, a model of fairness where scores reflect probabilities (calibration). In contrast, recent papers within the fairness literature have instead argued that scores should be equally distributed (demographic parity) or errors should not disproportionately affect certain races (equality of odds) [24].

Current techniques of fairness such as variable threshold rates by rates often fix one definition at the sake of the other two [24]. This is problematic from a legal perspective. As described in a recent working paper [5], the law students we have been working with have shown that a model that does not (approximately) meet all three definitions could be legally challenged and be considered unfair [39]. This is the main problem we consider in this thesis:

How can machine learning models be trained to approximately meet multiple definitions of fairness?

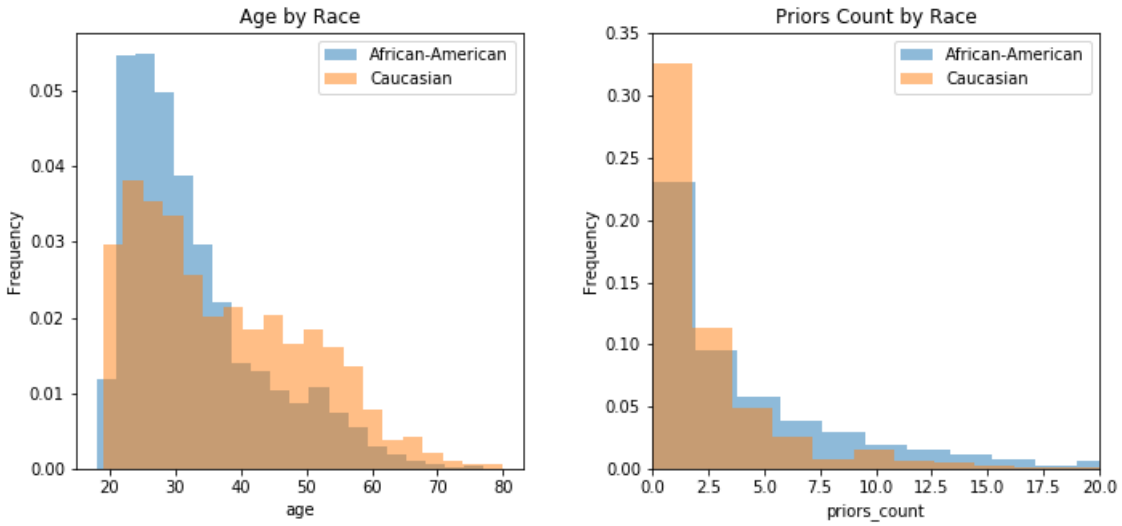
The main technical contribution of this paper is to test a set of methods on the proxy model build in chapter 2 and evaluate the impact on the three definitions of fairness.

For the first method of adversarial examples, consider a simplified model that only considers the number of priors and age of the defendant. Recent work has shown

[4]We proposed buying evaluations and re-training a model to match COMPAS based on survey answers. However, due to concerns about proprietary technology, this strategy was not approved by the IRB.

[5]In pre-print.

that a linear regression model of these two features is just as accurate as COMPAS ($\approx 67\%$) [17]. With a positive coefficient on priors and negative coefficient on age, younger individuals with more priors are given high risk scores. While this model does not directly consider a protected class such as race, it can still have discriminatory impacts. Figure 1.3 shows how African-American defendants are more likely to be younger and have a larger number of priors, leading to higher risk scores by this simple model.



(a) Age *(b) Priors Count*

Figure 1.3: Distribution of Age and Priors Between Races

In light of these distributional differences, we consider adversarial examples for the machine learning models. The example seen in figure 1.4 highlights how small changes in number of priors and age can lead to higher risk scores. Because we expect historical biases to play a role in these features, drastic changes based on a few number of priors and age go against our intuition of fairness. In chapter 4 we explore regularization strategies to enforce local consistency and diminish the impact of these adversarial examples. Results from this section show that this method is a first step in improving the fairness of risk scores.

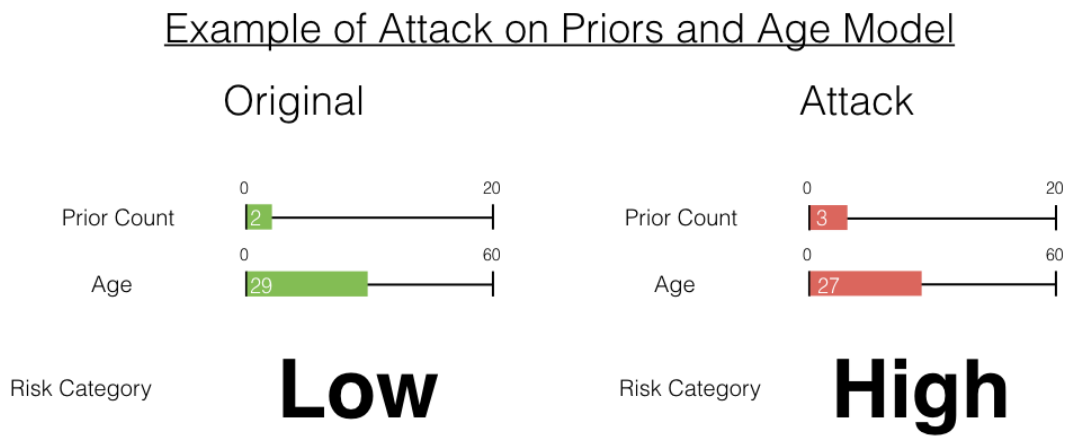


Figure 1.4: Adversarial Examples for Simplified Model of Priors and Age

The second method is introduced in 5. Instead of focusing on the inputted features, this method directly aims to satisfy the metrics for the various definitions of fairness. To do so, we introduce an adversarial network connected to the original network that attempts to predict race from outputted scores. In training, we maximize the loss of this adversary in order to remove any correlation between risk scores and protected categories such as race. This leads to a predictor that should not only be accurate but also approximately meets the definitions of fairness discussed.

In concluding, chapter 6 notes the success of both training strategies proposed on the proxy model. These models are significant steps in meeting the definitions of fairness proposed by the law students. We then ask whether these methods should be implemented in the real-world. Our first concern is that these methods showed success on the proxy model built, but we would need to analyze the impact on a full prediction algorithm such as COMPAS for enacting these correction methods. Outside a legal perspective, we also note contradictory literature within fairness that argues (1) against the removal of race or (2) that feedback loops may prevent these methods from being successful. In light of these papers, we propose directions for future research.

Also, this notion of fairness is only one of the legal arguments against the COMPAS model. Appendix B provides a more holistic view and describes the other arguments from the law student's paper.

Chapter 2

Reconstruction of COMPAS

2.1 ProPublica Dataset

To start our analysis of the COMPAS algorithm, we explore the dataset collected by ProPublica in 2016. The records comes from a collection of over 10,000 defendants and inmates in Broward County, Florida obtained through Freedom of Information Act requests. Features include publicly available information such demographic information, criminal history, and circumstances of all arrests. A list of all variables available for our analysis is included in figure 2.1. Better yet, the long time span of available data (over 2 years) allowed the researchers to measure true recidivism values based on whether or not a released inmate re-offended within the given time span [9, 37].

The only difference from the original ProPublica article used in this paper is that we further restrict to cases where the defendant is either African-American or Caucasian to provide a binary indicator for race and ensure there are enough trials to consider. This results in a total of 5278 cases to be analyzed from general COMPAS algorithms.

Features :	Age, Age Category, Priors Count, Juv Felonies, Juv Misdemeanors Days in Jail, Juv Other Count, Days before Screening Arrest Charge Degree, Sex, Race
Target Variables :	Decile Score, Risk Category
True Outcomes :	Two Year Recidivism

Figure 2.1: ProPublica Dataset Features

Unfortunately, the data available does not include responses to the 137-part COMPAS questionnaire, and model parameters are still unavailable. However, for our study of COMPAS, the dataset provides the decile scores that purport to predict the probability that an individual defendant will recidivate. When provided to judges, these scores are also often thresholded into risk categories (Low, Medium, High). Following ProPublica's analysis, we create a binary predictor for risk category by considering all decile scores under 5 as Low and grouping the Medium / High predictions together.

2.2 Feature Importance

In order to confirm the alleged discrimination noted earlier, we want to find the importance of various demographic and criminal history features on COMPAS predictions. To do so, we assume a linear model and fit a logistic regression model on the thresholded score (Low vs [Medium / High]) risk categories) using a subset of the features described above including age, priors, race, sex, crime degree, and the true two year recidivism outcome [30, 7]. Note that this is equivalent to running a generalized linear model (GLM) with a binomial noise model. The GLM has the added benefit of calculating the variance on the weights and determining statistical significance. The results for the general COMPAS model is seen in figure 2.2 below[1].

[1]See appendix A.1 for a discussion of the COMPAS model for violent recidivism

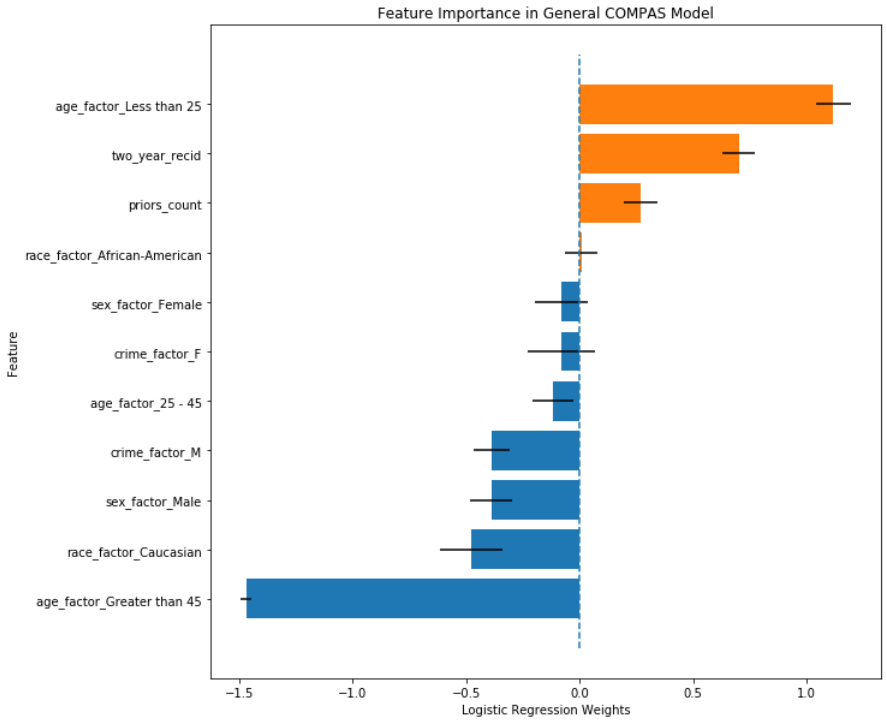


Figure 2.2: Feature Importance in Logistic Model of General COMPAS

What is interesting in the weights above is that the true two year recidivism is not the most important feature leading to a high score. Rather, the highest weight is given to the age of defendant being less than 25. Despite the distributional differences seen in 1, recent court cases have considered this feature to be probative and reasonable to include within such recidivism models. The same is true for the third most important feature of priors count.

Next, we turn our attention to the impact of features such as race and sex. The absolute magnitude of these weights is not as important as the relative difference between the two possible options. Note there is a discrepancy of almost 0.5 between the two races considered. This significant difference[2] provides more evidence for the discriminatory impact of COMPAS predictions.

It is important to note that this logistic model avoids many of the nonlinearities and covariance that are likely to exist in the data and incorporated within the true COMPAS model. However, it is a good first step in analyzing feature importance.

[2]Confirmed by the p-values in the GLM summary of 0.00 to 2 significant figures

2.3 Black-Box Reconstruction: *PROXY*

As mentioned in chapter 1, our goal in this thesis is to test a suite of a correction methods to machine learning models. However, at this point, we come face-to-face with the problem that we still do not have access to any model that is able to make predictions. Model parameters from COMPAS are still proprietary, and we also do not have any of the survey questions or answers that could be used to retrain a COMPAS model.

Instead, we turn towards a common technique for black-box model reconstruction: neural networks. Specifically, using only the features available within the ProPublica dataset, we create a new predictor of the COMPAS risk score. Neural networks are particularly well suited for this reconstruction tasks due to their flexible properties as universal function approximators. A sufficiently wide neural network can map any function up to arbitrary ϵ accuracy [25]. The non-linear dynamics between the hidden layers and output can better capture the covariation between available features and COMPAS's original model. This method is a generic and practical way to mimic what is happening within black-box models [35]. We refer to this reconstruction as *PROXY*.

It should be noted that as a reconstruction to the original model, this network studies how available features such as demographic information and criminal history play into COMPAS risk scores. However, it does not tell us anything about how the original COMPAS model operates. The inputs to these two are significantly different, the interactions between the variables in *PROXY* do not necessarily reflect COMPAS. In effect, it might be best to view *PROXY* as a toy model for criminal recidivism prediction that produces outputs similar to COMPAS. Even if it is not entirely realistic, the benefit of this model is now we have access to a model that we evaluate along with COMPAS in chapter 3 and then correct in chapters 4 and 5 [28].

2.4 Neural Network Details

There is limited literature on how to design a black-box reconstruction and what architectures should be used. A number of the papers indicate the network depends significantly on expected complexity of the original model and number of data points available. After a few rounds of testing, we settled on the following network details.

PROXY takes the 11 features directly from the ProPublica dataset[3]. The network

[3]The 11 features as described earlier in the chapter are Age, Age Category (2 variables), Priors Count, Juv Felony Count, Juv Misdemeanor Count, Days in Jail, Juv Other Count, Days before Screening Arrest, Sex, Charge Degree

consists of 1-hidden layer of 256 hidden units and uses a ReLU activation functions. The output layer is also linear with a sigmoid activation function. The resulting output is in the range [0,1], which once trained should reflect probabilities of recidivism. The output is trained with respect to the decile scores provided by COMPAS, using a binary cross entropy loss. Training is done on 80% of the available ProPublica data and utilizes the Adam optimizer with a learning rate of 0.001 for 600 epochs [28].

2.5 Comparison to COMPAS

In this section, we confirm that the $PROXY$ model achieves is similar to the COMPAS prediction. On the test set, the accuracy of $PROXY$ is 68.1% which is comparable to the accuracy of COMPAS of 64.1 %. Moreover, the proxy model predicts the same risk category as COMPAS in 74.2% of cases.

We also look at the differences between the COMPAS and $PROXY$ predictions. Figure 2.3 shows the residuals of the $PROXY$ reconstructions. From this graph, it is clear that $PROXY$ captures some of the basic trends of the COMPAS model, but for a number of defendants the risk scores from the proxy model is completely different. This highlights that while the $PROXY$ is a good reconstruction, it does not capture the dynamics of COMPAS and should be thought of as a toy model.

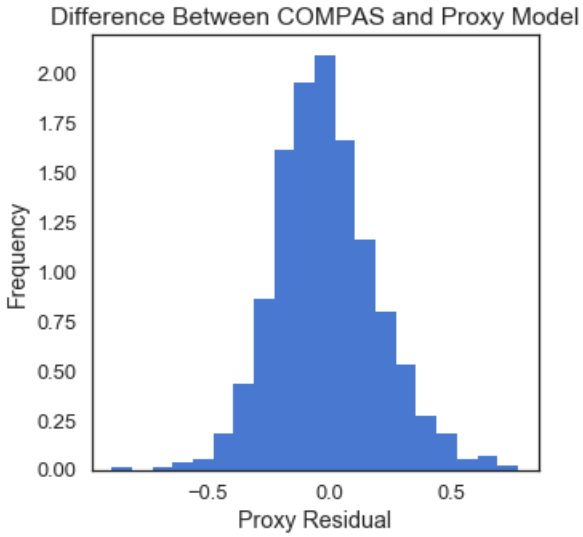


Figure 2.3: Difference between $PROXY$ and COMPAS

Chapter 3

Definitions of Fairness

3.1 Protected Classes and Fairness

The problem in creating models that are fair is that data alone cannot distinguish between meaningful patterns and historical biases[1]. To define when it is appropriate to use information on race, sex, or other such variables, we turn to the legal literature on protected classes. Specifically, federal law in the United State prevents discrimination against groups that have been historically marginalized. For example, the hallmark of anti-discrimination legislation is the Civil Rights Act of 1964 which protects individuals from differential treatment based on nationality, religion, and race. Additional protected classes have been added progressively over time.

As algorithms make their way into social domains, well-meaning practitioners first attempted to create fair models by removing details about the protected attributes described above. Unfortunately, fairness is not so simple. In 2012, Dwork et al. showed that such methods of 'fairness through unawareness' are ineffective[2]. The exclusion of a protected feature does not remove information or ensure that results are unbiased. For example, features such as zip code or profession could encode for information about race or sex respectively. Therefore, models can still have a differential impact through alternate variables [18]. So what does it mean for a model to be fair?

In this chapter, we explore three different mathematical definitions of fairness. COMPAS currently is trained to meet the definition of fairness known as calibration. How-

[1] In fact, the data that should or should not be used depends significantly on the exact application. In section 1.1, we saw a number of examples where the use of race or sex was likely unethical such as in employment decision. In contrast, consider the use cases within medicine. The very same features can be tremendously helpful in diagnosing diseases or providing appropriate care.

[2] This formalizes the distributional differences in features presented in chapter 1.

ever, recent literature has proposed countering definitions to minimize the discriminatory impact on different races, specifically demographic parity and equal opportunity[3]. We then discuss the inconsistencies between these models and how a newly proposed Equal Protection argument describes how a model that does not meet all 3 definitions could be legally challenged as unfair [39].

3.2 Supervised Learning Notation

Before diving into the legal framework, table 3.1 defines few pieces of notation considering the traditional supervised learning scenario. This notation will be used throughout this chapter to provide mathematical definitions for fairness.

Variable	Interpretation
X	collection of feature variables
Y	target variable, in this case true recidivism outcome
$\hat{Y}$	predicted outcome based on observed characteristics, $E[Y \mid X]$
A	binary variable for legally protected class such as race

Table 3.1: Supervised Learning Notation

3.3 Calibration and Sufficiency

When the COMPAS model was first designed, Northpointe's validations studies highlighted how model predictions were designed to reflect the probability of recidivism for any given individual. This is exactly the definition of calibration (by groups) within the machine learning context,

$$P(Y = 1 \mid \hat{Y} = \hat{y}, A = a) = \hat{y}, a \in \{0, 1\}$$

Outside the binary case, calibration generalizes to what is known as sufficiency. This condition argues that the protected attribute is included within the score to the extent that it provides predictive power, or more mathematically:

$$Y \perp\!\!\!\perp A \mid \hat{Y}$$

[3]It should be noted that there are many more definitions of fairness within the greater legal and computer science literature. At the Conference on Fairness, Accountability and Transparent (FAT ML), Arvind Narayanan highlighted over 21 definitions that have been proposed. In this thesis, we only consider the 3 mentioned above due to their widespread use and relevance to the legal argument.

Within the machine learning literature, calibration is often thought of as the default measure of fairness. Optimizing for accuracy or many common loss functions such as cross entropy loss will maximize the predictive power of the predictions and lead to these probabilities without any additional work. This has been noted for a number of problems such as income prediction [10].

As it relates to COMPAS, the states implementing these models and even ProPublica's article confirmed that scores were well-calibrated [15, 20]. In these studies, the metric used to measure calibration is the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), a statistic that we maintain in this paper. For our analysis of correction methods in later chapters, the baseline AUC-ROC for the COMPAS model and $PROXY$ are seen in the table below 3.2.

	AUC-ROC
COMPAS	0.707
$PROXY$	0.717

Table 3.2: Statistics for Calibration

3.4 Demographic Parity

As noted within fairness through unawareness, the assumption that race is included in data even when there is no overlap between the two datasets argues that

$$X \not\perp\!\!\!\perp A \rightarrow \hat{Y} \not\perp\!\!\!\perp A$$

Therefore, even under calibration we are likely to see disparate impacts between the races as our predicted target is not independent to our protected classes. Correcting for this leads immediately to our next definition of fairness.

Demographic parity[4] enforces that the distribution of scores for any protected classes is the same. This implies the independence of scores from protected class such that:

$$\hat{Y} \perp\!\!\!\perp A$$

For the COMPAS data, we analyze the scores from the ProPublica data. Figure 3.1 shows the differences in the distribution of scores for black and white defendants.

[4]Within the fairness literature, this definition is also referred to as statistical parity or disparate impact [10].

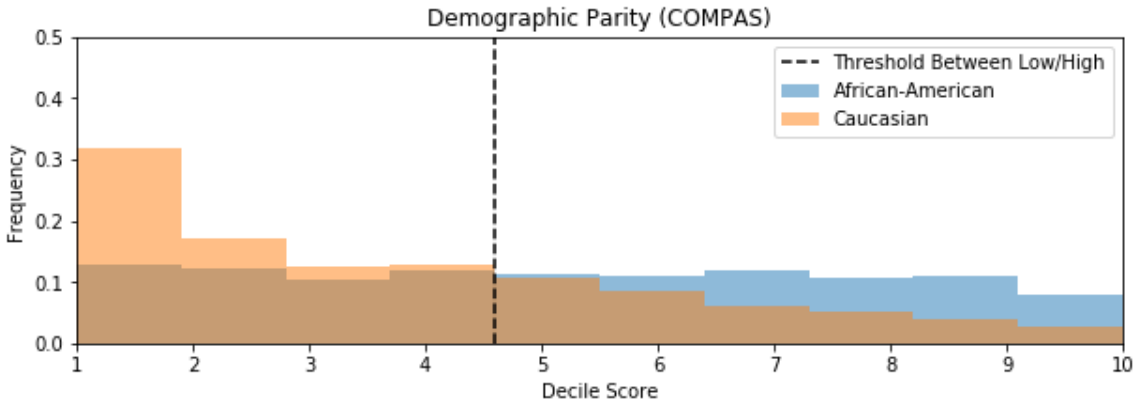


Figure 3.1: Demographic Parity in COMPAS

When the score function ($\hat{Y}$) is a binary output variable, this independence criterion can be further written as:

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

We consider the binary classifier from COMPAS of whether the defendant has a High risk score. Note, that this is different from the calibration $\hat{Y}$ as this requires thresholding the original probability or decile score produced from the model. For convenience, we have maintained this predictor notation in both cases, but it should be understood that there is this intermediate step. Table 3.3 highlights this difference in parity in the ProPublica data.

	% Low	% High
Caucasian	66.9	33.1
African-American	42.4	57.6

Table 3.3: Risk Category by Race

For the rest of this paper, we will use

$$|P(\hat{Y} = 1|A = 0) - P(\hat{Y} = 1|A = 1)| \tag{3.1}$$

as a metric for demographic parity[5]. For COMPAS and the *PROXY* model, the baselines for this definition of fairness are given in table 3.4

[5]Recent case studies of legal literature have proposed a 80% rule for determining discrimination. This could be interpreted as the metric proposed for parity being above 0.2. Instead of this additive error, another legal understanding for this rule is a multiplicative error of 0.8. Since our results are comparative, this is not as significant but good to keep in mind [10, 19].

	Demographic Parity
COMPAS	24.5%
PROXY	22.6%

Table 3.4: Statistics for Demographic Parity

3.5 Equality of Odds

The immediate concern with demographic parity is that the model does not depend on true outcomes. For example, consider a scenario where one protected class was guaranteed to recidivate and the other one would not. Equalizing the rate between these 2 groups would lead to significant loss [24].

Equality of odds addresses this problem by ensuring that $\hat{Y} = Y$ is always an acceptable answer. This fairness definition enforces that

$$\hat{Y} \perp\!\!\!\perp A \mid Y$$

which within the literature is also referred to as separation. This means that all correlation between the predictor and race must be justified by true outcome [10]. This is easier to understand in the binary case where equality of odds means that error rates must be equal, written as

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

The formula above can be understood as equalizing the true positive rate between races for $y = 1$ and false positive rate for $y = 0$. Equal opportunity is a slightly weaker fairness criteria and only focuses on this case where $y = 1$; this definition can be thought of as a way to equalize the chance that members of either race are given a high prediction conditional on their true outcome.

In regards to COMPAS, table 3.5 shows the relevant values for the equality of odds definition.

African-American		
	% Low	% High
Recidivated	28.5 (FNR)	71.5 (TPR)
Did Not Recidivate	57.7 (TNR)	42.3 (FPR)
Caucasian		
	% Low	% High
Recidivated	49.6 (FNR)	50.4 (TPR)
Did Not Recidivate	78.0 (TNR)	22.0 (FPR)

Table 3.5: Statistics for Equality of Odds

There are two particularly concerning statistics in the equality of odds for the COMPAS data. First, the false positive rate is almost twice as high for African-Americans, meaning African-Americans that do not recidivate are often given high scores. Also, the table indicates that high-risk Caucasians are given low scores up to 49.6% of the time. Clearly, COMPAS does not satisfy equality of odds fairness.

Throughout this paper, the metrics used for equality of odds will be the absolute difference in True Positive Rates (TPR) and False Positive Rates (FPR). These are expressed as:

$$|P(\hat{Y} = 1|A = 0, Y = 1) - P(\hat{Y} = 1|A = 1, Y = 1)| = \text{ TPR Difference}$$
$$|P(\hat{Y} = 1|A = 0, Y = 0) - P(\hat{Y} = 1|A = 1, Y = 0)| = \text{ FPR Difference}$$

The values for COMPAS and the $PROXY$ model were calculated and are presented in table 3.6.

	TPR Difference	FPR Difference
COMPAS	21.1%	20.3%
$PROXY$	22.6%	15.7%

Table 3.6: Statistics for Equality of Odds

3.6 Inconsistencies and Legal Impact

Given these very different definitions, one of the first questions is why not create a model that satisfies all 3? Sadly, recent literature has proven that these definitions are all pairwise inconsistent; there always exist cases where no two can be held at the same time. Theoretical counterexamples can be seen in [10] but are not worth discussing here. More practically, simple algorithms such a varying thresholds to enforce any one of the 3 definitions do so at the expense of the other two.

The fact that multiple definitions cannot be held at the same time is particularly concerning from a legal perspective. The law students we are working with have argued that under a *Batson v. Kentucky* perspective on Equal Protection, any of these three definitions could be used to argue that is a model is unfair. Therefore, a fix to COMPAS would have to satisfy all 3 definitions [39].

In the rest of this paper, we consider more complex methods that attempt to balance and approximately meet these definitions of fairness. We start with the $PROXY$, which like COMPAS that has high-calibration but poor parity and equality of odds. Generally, within the methods described in the next part, we see that giving up a small amount of calibration allows us to better meet the other definitions of fairness as measured by the metrics discussed above and relayed in table 3.7.

Definition	Measure
Calibration	AUC-ROC on Unthresholded $\hat{Y}$
Demographic Parity	$\lvert P(\hat{Y}=1\lvert A=0) - P(\hat{Y}=1\lvert A=1)\rvert$
Equal Opportunity	$\lvert P(\hat{Y}=1\lvert A=0, Y=1) - P(\hat{Y}=1\lvert A=1, Y=1)\rvert$

Table 3.7: Metrics Used to Evaluate Definitions of Fairness

Part II

Methods and Results

Chapter 4

Adversarial Examples

4.1 Motivation and Literature Review

Adversarial examples in deep learning were first introduced by Szegedy et al. in 2013 [40]. The researchers noticed that neural networks for computer vision were particularly vulnerable to minor changes in inputs. Strategic noise could be used to lead to dramatically different predictions, even though the images appear the same to most observers. [22]. While initial examples from literature focused primarily on high-dimensional computer vision scenarios, recent research has brought the same kind of attacks to other fields in machine learning. For example, Sarkar et al. discuss the problem of adversarial examples within credit scoring [38].

In the application to fairness, in chapter 1 we explored how the distributions of features such as priors and age are different between groups of protected classes. This is a common trend between the variables in the ProPublica dataset. The problem is that neural networks or complex machine learning algorithms may assign drastically different scorers based on the minor perturbations that separate classes. This is highlighted in an adversarial attack on the $PROXY$ model where the risk score goes from low to high solely based on small changes to a few features, as seen in 4.1.

Adversarial defense methods attempt to minimize the impact of such strategic noise and enforce local consistency. For example, one such defense method is known as adversarial training and has been shown to be an effective regularizer and have beneficial impacts on robustness [22, 21]. These defense methods revolve around constructing and including adversarial examples during model training. The next section dives into common adversarial attacks, and then we return to how these adversarial defenses operate and perform in the context of fairness.

Figure 4.1: Adversarial Attack Example on PROXY model

4.2 Adversarial Examples Notation

Adversarial methods operate on continuous variables, creating minor changes under specific budget constraints. In this paper, we focus on gradient-based attacks that target most if not all of the features and perturb by a small amount. [1]

In the next few sections, we explain a few common adversarial attacks where all are based on the following notation.

$\mathbf{x}$: original features
θ : model parameters
y : label

[1]The other class of attacks is known as saliency map attacks. These target a few number of characteristics but allows for a greater degree of change. This corresponds less to our problem since the discrepancies seen between races are often small and over a large number of features. Early results for using this methods were not successful.

$$
\begin{aligned}
L(\mathbf{x}, \theta, \mathbf{y}) &: \text{ loss function} \\
\mathbf{x}^* &: \text{ perturbed features} \\
f(\mathbf{x}) &: \text{ classification probabilities}
\end{aligned}
$$

4.3 Gradient-Based Attacks

Two of the most common gradient based attacks within recent literature have been the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (FGSM)

FGSM was one of the earliest proposed attacks and involves adding the sign of the gradient of the loss function to particular input vectors. By moving in a direction similar to the gradient, the loss function should be significantly increased. The sign operator however ensures that the degree of perturbation is fairly uniform across the elements of the feature vector. This can be written as:

$$\mathbf{x}^* = \mathbf{x} + \epsilon \times \text{sign}(\nabla_{\mathbf{x}} \mathbf{L}(\mathbf{x}, \theta, \mathbf{y}))$$

where ϵ is a hyperparameter that controls the degree of noise [22].

PGD similar incorporates gradient ascent but without the sign operator. Instead, this method introduces a procedural construction. The original example is initially randomly perturbed to be within a ϵ of the original image. Then, the k small steps of gradient ascent are taken with clipping to ensure that the output image is no more than ϵ away from the original.

$$
\begin{aligned}
\mathbf{x_0^*} &= \text{ Random Initialization within } B(\mathbf{x}, \epsilon) \\
\mathbf{x_i^*} &= \text{ clip}_\epsilon(\mathbf{x_{i-1}} + \alpha \times (\nabla_{\mathbf{x}} \mathbf{L}(\mathbf{x}, \theta, \mathbf{y})), \mathbf{i} \in \mathbf{1}, \mathbf{2}, \ldots, \mathbf{k} \\
\mathbf{x}^* &= \mathbf{x_k}
\end{aligned}
$$

k, α here are secondary hyperparameters that control the number of gradient ascent steps and size of these steps respectively. Due to the random restarts and multiple steps, Projected Gradient Descent has been shown to be one of the stronger attack methods [33].

4.4 Adversarial Training

Adversarial training was proposed in the initial paper on adversarial examples and remains the go-to method for countering attacks and enforcing local consistency. The method suggests creating and adding adversarial examples to the training mini-batch. This is particularly effective at minimizing the impact of minor perturbations and lead

to general improvements in robustness [22, 21].

Algorithm 1 further delineates the specifics of adversarial training and provides the details of our specific implementation such as split size. Note that in the algorithm the perturbation is chosen beforehand, without knowledge of which is most significant. In the experimental results section, we refer to these models as $FGSM$ and PGD respectively.

Data: mini-batch during neural network training
for mini-batch X in training of size N **do**
  X_{adv}, $X_{clean} = X[: \frac{N}{4}]$, $X[\frac{N}{4} :]$;
  X^*_{adv} = Adversarially Perturb X_{adv};
  X^* = concatenate X^*_{adv} and X_{clean};
  Calculate $L(X^*, \theta, \mathbf{y})$;
  Update parameters θ using Adam optimizer;
end
Result: adversarially trained model

Algorithm 1: Adversarial Training [22, 21]

4.5 Experimental Results

In this section, we provide a brief comparison between the COMPAS, $PROXY$, $FGSM$ and PGD tables. We first consider the effect of the adversarial training on the predictive power of these networks in table 4.1.

Model	Calibration Score (AUC-ROC)
COMPAS	0.707
$PROXY$	0.717
$FGSM$	0.710
PGD	0.702

Table 4.1: Calibration Scores Between Models

Clearly, there has been some sacrifice of calibration for the adversarially trained models but the difference is not large. Next, we look at the parity plots created for these models to get the first signs of improvement over the $PROXY$ model:

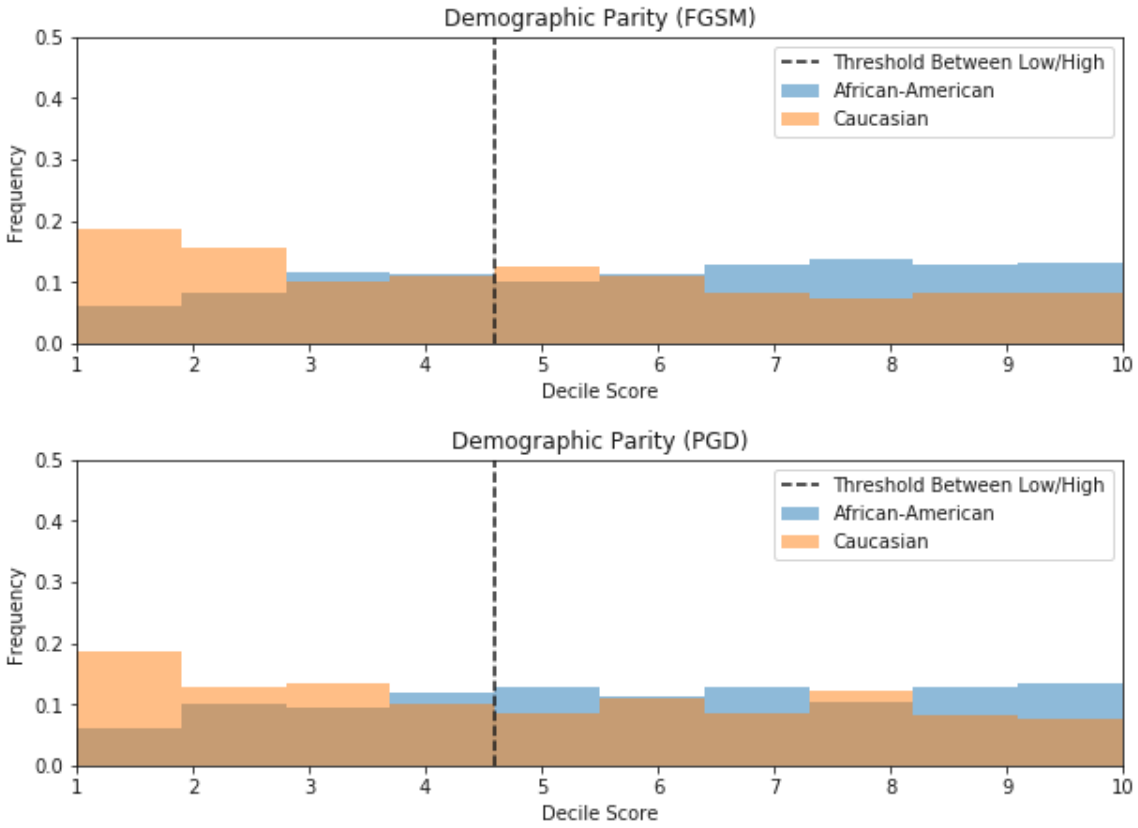


Figure 4.2: Parity Plots of FGSM (top) and PGD (bottom) trained models.

Although the score distributions are slightly different based on race, these graphs implicitly show some improvement over the baseline $PROXY$ model. Table 4.2 confirms this through our metric for parity.

Model	Parity Score
$COMPAS$	24.5%
$PROXY$	22.6%
$FGSM$	17.9%
PGD	10.4%

Table 4.2: Parity Scores Between Models

Both adversarially trained models show improvement over baselines results. In fact, a hyper-parameter optimized version of the $FGSM$ and PGD attacks gets to a parity of as low as 8% but leads to a much larger decrease in calibration. Finally, we evaluate the equality of odds metrics proposed in the chapter 3, the difference in true positive rates and false positive rates between races. These are explored in figure 4.3.

Model	True Positive Rate Difference	False Positive Rate Difference
COMPAS	21.1%	20.3%
PROXY	22.6%	15.7%
FGSM	21.5%	7.8%
PGD	12.9%	1.6%

Table 4.3: Equality of Odds Scores Between Models

4.6 Conclusion

As mentioned in the introduction to this section, adversarial examples are not specifically designed to help meet various fairness metrics. Rather, the method of adversarial training provides an effective regularizer that enforces a notion of local consistency such that minor changes in inputs do not lead to large changes in predictions. In the context of fairness, this in turn reduces the impact of the small differences between feature variables between races.

From the results section, adversarial examples performed remarkably well for not directly targeting the fairness metrics. The PGD-trained model cut the parity score and TPR difference by almost half, and the final FPR rate was 10 times less than the original proxy model. The only loss of using this method was a slight decrease in calibration. Within the adversarial literature, PGD-trained model are the most effective at minimizing the effect of adversarial examples, so the corresponding improvement in results is another sign that this strategy of gradient-based adversarial training is an effective way to remove differences between races.

A step in the right direction of meeting multiple definitions of fairness, training with adversarial examples provided significant benefits to the *PROXY* model.

Chapter 5

Adversarial Networks

5.1 Motivation and Literature Review

Adversarial machine learning has arisen as a set of techniques to remove domain correlations within data. This technique usually refers to the use of multiple neural networks, each with its own competing loss function. Since we have multiple definitions of fairness, this idea of competing models seems particularly promising to find a balance between calibration, equality of odds, and demographic parity.

Past research has attempted to bring similar methods to training fair models. For example, the early work of Beutel et al. enforced demographic parity between groups by sharing hidden layers between groups [12]. More recent papers have attempted to build off the popular topic of General Adversarial Networks. In such models, one neural network attempts to generate fake samples and the other discriminates between real and fake samples. A reverse gradient is applied on the discriminator / predictor for the protected classes, ensuring that the features and predictions do not encode for race [42].

In this chapter, we replicate and extend the contributions of Wadsworth et al. [41]. The model recreates the two-network game common to adversarial literature. The predictor is concerned with maximizing predictive power and making scores well-calibrated. The second network, the adversary, has a loss function designed to remove information about the protected class. This two-network game is therefore specifically designed such that the resulting predictions balance multiple definitions of fairness.

5.2 Network Setup

As mentioned in the previous section, the general setup for these adversarial networks starts with a predictor network N_N that generates predictions $\hat{Y}$. When trained with the usual loss functions such as MSE or cross entropy loss, this network N_N will maximize the predictive power and correspondingly the calibration of the outputted scores.

In contrast, the adversarial neural network N_A is designed to enforce one of the other definitions of fairness. Effectively, this network N_A want to remove information about the protected class from the outputted score distribution. Figure 5.1 delineates the general structure of these two network N_A and N_N.

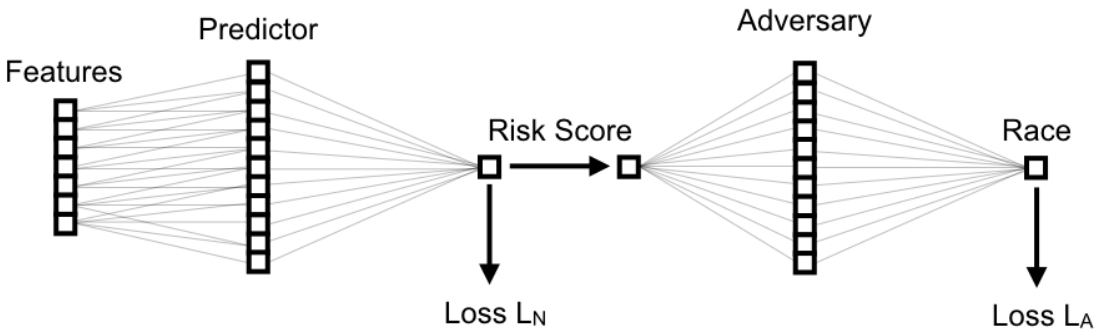


Figure 5.1: Diagram of Neural Network [41]

The best part of this general layout is that the adversary can be designed to meet various definitions of fairness. Below, we describe three possible setups for N_A:

1. **Demographic parity** ($PARITY$)

   The goal of demographic parity is to ensure a similarity between the distribution of scores between various protected classes. Therefore, to minimize the discrepancy, we design network N_A to predict the protected class A based only on the output recidivism score ($\hat{Y}$). This loss function is propagated backwards through model N_N to create a a 'fair' model. Resulting scores should not only be well calibrated but also be distributionally similar for all protected classes.

2. **Equality of Odds / Equal Opportunity** ($EQUAL$)

   In the network N_A for demographic parity, the network only predicts race from one's score. In equality of odds, we are not concerned about the prediction itself but whether that score was indicative of the true outcome. To modify network N_A and enforce equality of odds, the inputted metric will not be the absolute

score. Rather, N_A will take in the difference between the predicted outcome and the true outcome $\hat{Y} - Y$.

3. **Multiple Definitions of Fairness** ($MULT$)

   Finally, we consider the case of providing both the predicted score and true score to the network N_A. Because predicted score is one of the inputs, we expect this model to meet demographic parity. However, the non-linear dynamics of network N_A should also find the equality of odds metrics and minimize the discrepancy if it is not met. Note, this is the model described in Wadsworth et al. and Zhang et al. [41, 42].

5.3 Experiments

The experimental setup maintains the black box reconstruction as described in chapter 2. N_N is equivalent to $PROXY$ which is a 1-layer neural network with 256 units and ReLU nonlinearities. This should allow for consistent results when compared to estimates from both chapters 3 and 4. As before, this network is trained using and binary cross entropy loss to the decile score provided by COMPAS. Let this loss be denoted as L_N.

Within the greater computer science literature, there is little guidance on the size of the adversarial network. Similar to N_N, we let N_A have 1-hidden layer of 32 hidden units and a linear output layer with sigmoid activation function. This network is evaluated on an binary cross entropy loss in comparison to the race feature. Let this loss be denoted as L_A, where we want to maximize this loss to correspond to the removal of class information.

The general approach to correcting a black-box model is to start by pre-training. This involves optimizing N_N on L_N for 600 epoch, meaning that this model will start as equivalent to the previously-defined $PROXY$ model. A similar method is used to pre-train N_A, although it is only done for 10 epochs due to simplicity of the model.

These two networks are then connected, and we train for another 100 epochs. Here, the loss function used for backpropagation of both networks is a weighted combination of both L_N and L_A as described above:

$$\mathcal{L} = L_N - \alpha L_A$$

α here is a positive hyper-parameter that corresponds to the balance between the how important calibration is (improving L_N) and how much the adversary should enforce the other definitions of fairness. As found in [41] for the income prediction problem, we use an α of 30. This was heuristically confirmed to be strong enough to lead to

the desired decreased in parity and equal opportunity but not too strong to lead to mode collapse on the recidivism prediction problem.

5.4 Results

As described in chapter 3, we refer back to the various definitions of fairness and related metrics to evaluate the effectiveness of these parameters. The definition of calibration simply corresponds to L_N as defined in the previous section. Table 5.1 showcases how enforcing any of these alternate definitions of fairness reduces the predictive power of the network.

Model	Calibration Score
COMPAS	0.707
$PROXY$	0.717
$PARITY$	0.700
$EQUAL$	0.722
$MULT$	0.694

Table 5.1: Calibration Scores Between Models

While this decrease was expected, the benefits of the model will be seen in the improvements in parity and equal opportunity. For parity, table 5.2 first showcases the benefit of the model compared to the baseline $PROXY$.

Model	Parity Score
COMPAS	24.5%
$PROXY$	22.6%
$PARITY$	1.2%
$EQUAL$	5.5%
$MULT$	0.3%

Table 5.2: Parity Scores Between Models

Figure 5.2 goes further to showcase that $PARITY$ meets its intended function of making the two score distributions very similar. It is also worth noting that $MULT$ also shows very similar improvements as seen in figure 5.3.

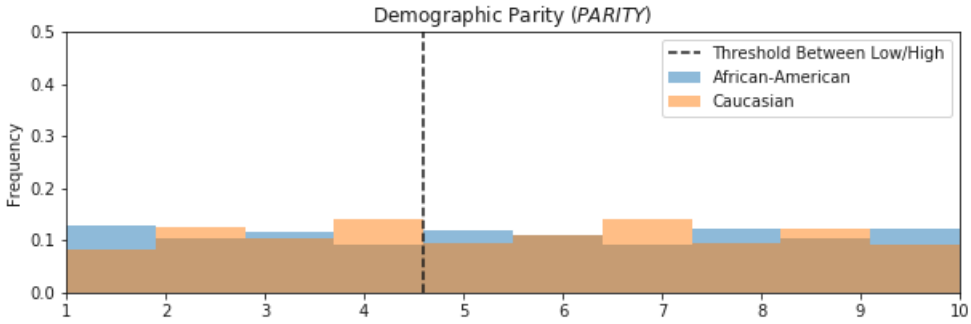


Figure 5.2: Parity Plot of PARITY Model

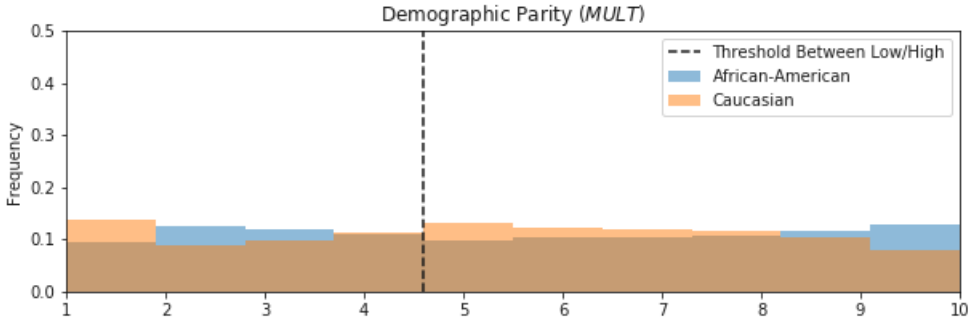


Figure 5.3: Parity Plot of MULT Model

Finally, moving on to equality of odds, the metrics is the comparison of the True Positive Rate (TPR) and False Positive Rate (FPR) of the various models. Table 5.3 demonstrates the effectiveness of these models in comparison to the COMPAS model.

Model	True Positive Rate Difference	False Positive Rate Difference
COMPAS	21.1%	20.3%
PROXY	22.6%	15.7%
PARITY	5.4%	4.1%
EQUAL	2.7%	0.7%
MULT	0.8%	0.3%

Table 5.3: Equality of Odds Scores Between Models

Through the usage of adversarial training for both demographic parity and equality of odds shows significant improvement in regards to the relevant fairness definitions discussed throughout this thesis.

5.5 Beyond COMPAS: Training from Recidivism

One element that was not emphasized in the previous sections is that we were trying to fix our reconstruction of $PROXY$. However, given the outcomes accessible within the ProPublica dataset, we can also reconstruct a model that predicts the true recidivism outcomes based on this demographic information. By removing the step of training on COMPAS, this removes the general patterns from the 137-question survey but could lead to better calibrated scores.

We again copy the structure of N_N and N_A. The only difference is that the output and loss function of N_N is based on the true recidivism outcome. The pure accuracy of this model starts at 65.2% which is slightly better than the COMPAS model on this particular test set. However, similar to COMPAS, we run into the problem that this model leads to a distinct disparity in score distributions for various races. This is seen in the top graph of figure 5.4.

Again, we run through the process of adding the adversarial model. And train for 100 epochs. The problem with this model is that since it is not based off of the COMPAS baselines, there are no comparable values or metrics to calculate. Any threshold to set would be arbitrary and not generally informative of the performance of the model in comparison to COMPAS. What we can do however is take this model and visualize the performance of the parity model as demonstrated in figure 5.4. This demonstrates how the methods introduced in this chapter are not specific to the COMPAS model or black box reconstructions but in effect can be used on any arbitrarily defined supervised learning problem.

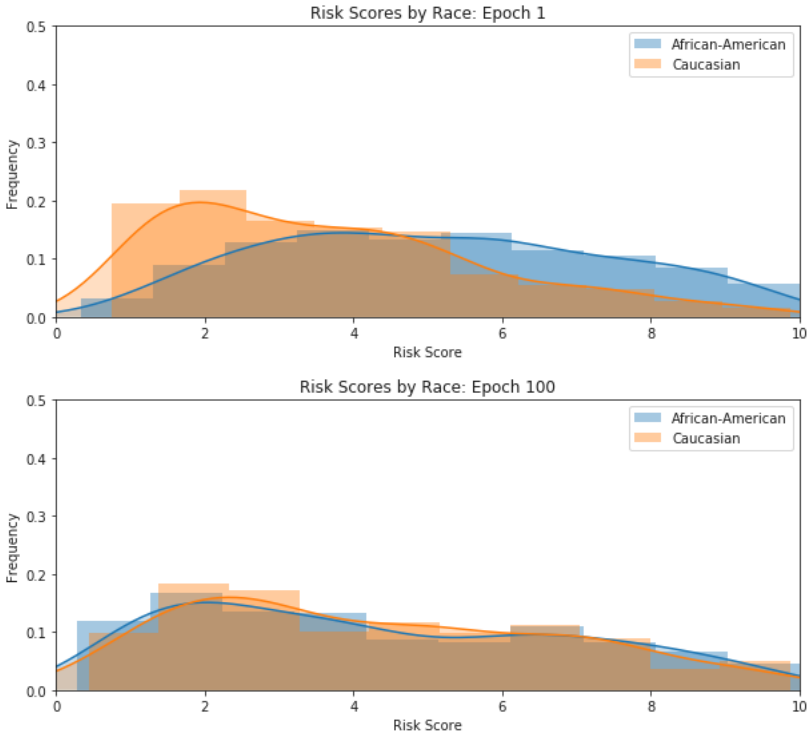


Figure 5.4: Parity Before and After Adversarial Training

5.6 Conclusion

As presented in this chapter, adversarial networks can clearly be used to improve various fairness metrics for arbitrary machine learning models. The two player game between the model leads to a balance between calibration from the predictor and the other definitions of fairness as defined by the choice of adversary.

For example, the $PARITY$ model worked very well at reducing the demographic parity metric from 22.6% to 1.2%, and the $EQUAL$ model improved its corresponding metrics as well. Of all the corrections to $PROXY$ proposed so far, $MULT$ led to the best parity and equality of odds metrics, but at the same time did have the lowest calibration. This is a general trend worth noting. In order to increase the fairness metrics of parity and equal opportunity, most models did so at the expense of slight amounts of calibration.

With respect to the legal framework, this network goes farther in approximating multiple definitions of fairness on $PROXY$. Furthermore, 5.5 highlighted how this method is generic and can be used to train models from scratch that meet multiple definitions of fairness instead of simply correcting a previous model.

Part III

Conclusions and Future Work

Chapter 6

Conclusion

6.1 Interpretation of Results

In this thesis, we attempted to bridge the gap between traditional computer science discussions of fairness and legal perspectives. Collaborating with researchers at Harvard Law School, we started with a legal framework of Equal Protection, which argues that only meeting one mathematical definition of fairness is not sufficient. Rather, all three definitions of calibration, demographic parity, and equal opportunity would need to be approximately satisfied to prevent legal challenges. This is especially true in the case of COMPAS and criminal recidivism prediction.

The methods explored in chapters 4 and 5 tested whether two different algorithmic techniques could be used to satisfy this legal definition of fairness. The first approach suggested the use of adversarial examples and defenses to enforce local consistency and prevent minor discrepancies between races from leading to large differences in predictions. Adversarially-trained models led to slight improvements in both parity and equality opportunity in comparison to the original $PROXY$ model, a good first step. However, with parity scores still over 10%, this model still left room for improvement.

Next, we explored the use of adversarial networks. Specifically, we consider the competition between two networks, a predictor aimed at increasing the predictive power of risk scores and an adversary designed to reduce the impact of protected classes such as race. The models designed in this section were extremely effective at balancing these definitions of fairness. Our final $MULT$ model had parity and equality of odds metrics all less than 1%, for only a slight decrease in calibration. Coming the closest to meeting the three definitions of fairness, the adversarial network method is promising as a way to develop models that satisfy legal constraints.

This leads us to the fundamental question:

> Should the two correction methods proposed in this thesis be used in real-world settings to make models legally fair?

Although supported by the results, we face a fundamental limitation in arguing for the use of adversarial examples or adversarial networks. Remember, the tests within this thesis are all on the toy $PROXY$ model. It uses different features and dynamics from the original $COMPAS$ algorithm, and there is no guarantee that the same methods would work as well in the real-world setting. Future work hopes to test these methods on other fairness problems and evaluate if this balance of fairness definitions is maintained.

Outside the legal perspective, it is also important to consider the broader ethical implications of these methods. Recent fairness literature has brought up major points of criticism about enforcing demographic parity or equal opportunity. We explore two of these concerns in the next sections.

6.2 Inclusion of Race

The inclusion of race has been a fundamental issue within this thesis. Parity models that are created by methods such as adversarial examples and adversarial networks effectively attempt to minimize the impact of race on risk scores.

In a recent paper, Kleinberg and Mullainathan argue that there is a fundamental trade-off between fairness and efficiency [29]. In order to create a model that is fair, often information about the individuals must be lost, and a simpler model than what would have otherwise been used is created. As we have already explored, this will lead to a decrease in performance of the model (ex: in terms of accuracy or calibration) to enforce various definitions of fairness. The authors take this a step further and show that using a simpler model will also decrease equity of the prediction, not necessarily of the protected group but of some other class.

The paper cites the example of college admissions. One recent proposal has been to eliminate the essay from admissions process in favor of only looking at test scores and recommendations. This method attempts to accommodate for certain groups having greater access to essay-writing resources. Although purported to increase the fairness of the model, this proposal disproportionately hurts various groups of disadvantaged students who could no longer cannot show excellence through an essay.

The machinery created in the paper is generic and explains that any fairness definition will complicate the issue of equity for the model. While the protected group may be

helped, some other group will face a greater problem of inequity. This is not significant from a legal standpoint, but from a greater ethical perspective, is this a fair treatment? Are we able to discover and define all protected classes that deserve to be equalized? By enforcing the definitions of parity or equal opportunity on COMPAS, we may otherwise be harming some group that is not explicitly defined as protected.

6.3 Temporal Effects and Delayed Fairness

Also, real-world decision are rarely made in isolation. If there was no temporal effect, the simple approaches of demographic parity and equal opportunity make sense in light of the discussed legal framework. This argument is supported by recent case studies such as Ensign et al. which have shown that predictive policing algorithms that are initially unfair can lead to a racial bias and focus on certain neighborhoods, even when the crime rates are equal [32].

However, delayed fairness is perhaps one of the most counter-intuitive results within fairness literature. As presented in Liu et al., delayed fairness argues that the temporal effects of enforcing decisions such as demographic parity or equal opportunity can lead to either positive or negative outcomes over time [31]. The paper focuses on the case of credit scoring and notes that defaulting on a loan is not only detrimental to the issuing bank but also to the individual as well who faces higher interests rates in future loans. An algorithm that meets the definitions of fairness of parity and equal opportunity is artificially inflating the scores of those individuals that may have a higher chance of defaulting. If these individuals do fail to repay the loans, they may be trapped in a cycle of lower credit scores. Similar logic is shown for advertising and college admissions and highlights that improvement over time is not guaranteed by enforcing demographic parity or equal opportunity. Results are supported by a case study on FICO credit scores, and the paper even shows how models that only use these alternate definitions of fairness as regularizers are still impacted by the temporal effects of delayed fairness.

In relation to the models described in this thesis, the results from delayed fairness argue that care should be taken to consider the temporal effects for implementing such models. In different scenarios, these changes may cause relative improvement or relative harm compared to baseline calibration-based / utility-maximizing models.

6.4 Future Work

Clearly, there are a number of contrasting perspective on whether or not fairness definitions such as demographic parity and equal opportunity should be enforced. This

confusion allows for a number of opportunities for future work. Following Kleinberg's logic directly, one could work to find the groups that become disadvantage based enforcing these fairness definitions. From Liu's work, it would be interesting to simulate outcome curves for COMPAS scores based on various definitions of fairness. This would be a pseudo-experiment for the temporal effects of the model; however, the lack of data availability from multiple time spans prevents such analysis.

More in line with this thesis, we are also interested in determining the applicability of the proposed methods to applications of machine learning outside criminal recidivism prediction. For example, Wadsworth et al. has looked deeper into income prediction, and Sarkar et al. worked on grading loans [41, 38]. Many other applications worth exploring exist.

Finally, the methods discussed throughout this thesis attempt to deal with a fundamental problem in machine learning: historical prejudices within data. As long as such sources are relied on for training, there is a major concern that the resulting models may encode for discriminatory predictions. One of the most promising directions within fairness literature to overcome this effect is causal inference. By asking what leads a defendant to re-offend and using external information about criminality, these models may be able to overlook historical biases and focus on the true causes of recidivism. For future work, we are interested in applying recently proposed causal models such as counterfactual regression to datasets such as ProPublica's[1].

[1]However, without the counterfactual data, predictions from these models may be difficult if not impossible to validate.

Appendices

Appendix A

Additional Analysis and Figures

A.1 Violent Recidivism COMPAS Model

COMPAS also provides a model that predicts the probability of a violent crime being committed to help inform judges if the defendant is a threat to others around him. From the ProPublica dataset, 3377 rows also had predictions for violent recidivism. From ProPublica's analysis, this dataset is similar to the general model and even features less of a racial disparity. This is confirmed in our basic logistic model as presented below which is very similar to the one seen for the general model in A.1. Due to these similarities, this model is not heavily discussed in this paper.

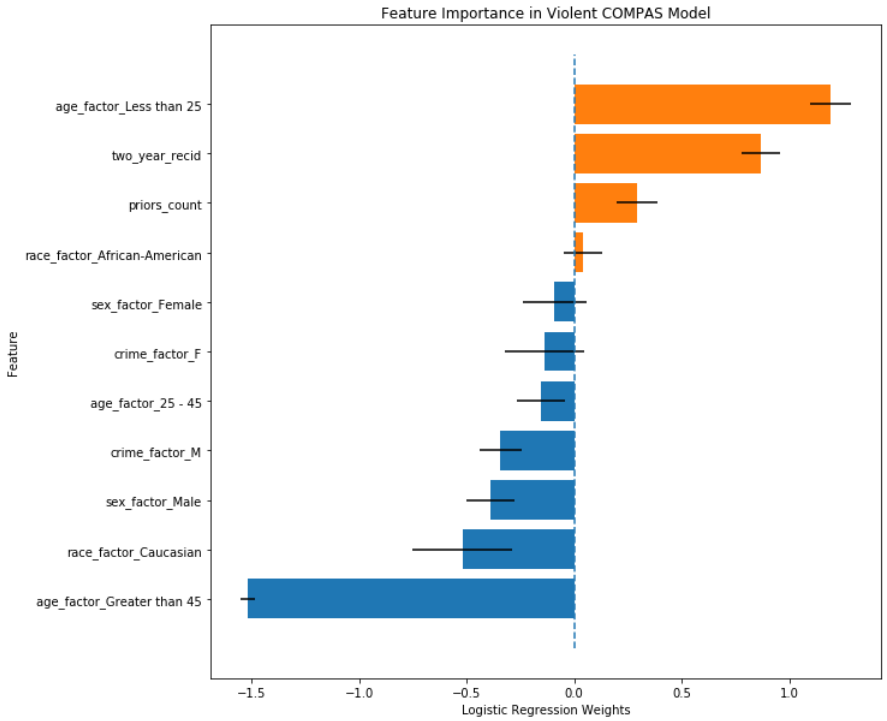


Figure A.1: Feature Importance in Logistic Model of Violent COMPAS

Appendix B

Ethical and Legal Implications

B.1 Introduction

The ideas presented in this thesis are a technical extension of on-going work with researchers at Harvard Law School working on the legality of using algorithms such as COMPAS within the justice system. Throughout this thesis, we have focused on the problem of disparate impact and its relation to the legal arguments of fairness from a Equal Protection perspective. In this chapter, we summarize the other possible legal challenges to machine learning models such as COMPAS as presented by the working paper by the law students we have been working with [39]. We start with the current legal framework and then explore Due Process concerns.

B.2 Current Legal Standing

The use of algorithms within the justice system is not new, but algorithms such as COMPAS have received recent criticism for the discriminatory impact that was highlighted by the investigative journalists at ProPublica. The differences in the scores led to challenges in various states where the recidivism prediction tools are used. For example, the Wisconsin state Supreme Court heard *Loomis v. Wisconsin* about the use of such predictions within criminal sentencing [6, 2]. The decision in that case remains the bulwark of support for these algorithms and concludes that such predictions (even if biased) can be used if they are one of a number of factors considered within a bail or sentencing decision.

B.3 Due Process

Due process as defined under the 14th amendment to the Constitution and more recent legal literature argues that an individual has the right to review and counter the arguments made against them in court. Algorithms are currently being challenged on two front: black-box nature and explainability.

B.3.i Black-box Nature

As mentioned in chapters 1 and 2, one of the predominant concerns with the COMPAS dataset and subsequent analysis is its proprietary nature. Created by a for-profit company, the model parameters are not released but rather the evaluation is licensed to the states that use the tool. This means that while the questions are available for study, the way the algorithm accumulates these answers into producing a score is not. Methods used could be as simple as linear regression to non-parametric approaches such as k-Nearest Neighbors. Without a means of analyzing this tool, defendants are left staring at a score with no idea how it was created or any way to challenge the model itself [39].

Black-box reconstruction methods are a first step in tackling this problem. The logistic regression was probably the simplest tool that could be used to test the importance of various features in COMPAS. More work could be done to analyze the neural network reconstruction as well, but there are few guarantees that these methods find the same functionality as COMPAS and therefore are significantly weaker than a simple release of the model parameters.

B.3.ii Explainability

The next problem with many machine learning methods is their lack of explainability. The question is whether the average person contesting their score within court would be able to understand how their answers are effecting the end score given by the recidivism prediction algorithm. On one end of the spectrum, we have linear or logistic regression algorithms. These can be fully described as the sum of weights on question answers. Tools such as ORAS (the recidivism prediction tool used in Ohio) can be thought to fall within this category, with predefined integer weights. Such a model would be easily challenged as one could argue for inappropriate weights or that a factor did not effect their probability of recidivism.

In contrast, with black-box models such as COMPAS, a concern is that the underlying algorithm is much more complex. For example, the underlying prediction function could come from a neural network or non-parametric models such as k-Nearest Neigh-

bors. Estimating the impact of question answers and their validity would be nearly impossible in these regimes, within the scope of a single court case. This is evidenced by the years of research put into understanding neural network predictions in the field of computer vision. Without this element of explainability, due process is similarly not fulfilled [39].

How could this requirement be fulfilled? There is no single method that has been deemed explainable, but linear / logistic regression probably remains the gold standard of machine learning algorithms. However, we have also explored two newer methods that seem to make some progress in this decision making process. First, a recent paper on Certifiably Optimal Rule Lists (CORELS) offers a way to make decision tree where there is a balance between the number of parameters and the prediction accuracy. This leaves a simple if-then set of rules that is common in many other legal domains and fits this thesis author's interpretation of explainability [8]. A second approach is inspired by FICO credit scores. While the algorithm is fairly complex, the report includes the most salient features that lead to the results, such as missed payment or lack of a credit history. By providing these scores for criminal recidivism prediction, judges and defendants would be able to understand how predictions arise and making individual decisions about the validity.

Bibliography

[1] Algorithms in the criminal justice system. Algorithmic Transparency.

[2] State v. loomis. volume 130.

[3] Risk assessment. 2011.

[4] Compas risk & need assessment system. 2012.

[5] Practitioner's guide to compas core. 2015.

[6] State v. loomis. Jul 2016.

[7] J. Adebayo. FairML: Auditing Black-Box Predictive Models. GitHub, 2017. https://github.com/adebayoj/fairml.

[8] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, and C. Rudin. Learning certifiably optimal rule lists. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 35–44. ACM, 2017.

[9] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. ProPublica, May, 23, 2016.

[10] S. Barocas, M. Hardt, and A. Narayanan. Fairness and Machine Learning. fairmlbook.org, 2018. http://www.fairmlbook.org.

[11] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943, 2018.

[12] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. arXiv preprint arXiv:1707.00075, 2017.

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

[14] R. M. Dawes, D. Faust, and P. E. Meehl. Clinical versus actuarial judgment. Science, 243(4899):1668–1674, 1989.

[15] W. Dieterich, C. Mendoza, and T. Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. Northpoint Inc, 2016.

[16] W. Dobbie, J. Goldin, and C. S. Yang. The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. American Economic Review, 108(2):201–40, 2018.

[17] J. Dressel and H. Farid. The accuracy, fairness, and limits of predicting recidivism. Science advances, 4(1), 2018.

[18] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference, pages 214–226. ACM, 2012.

[19] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 259–268. ACM, 2015.

[20] A. W. Flores, K. Bechtel, and C. T. Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. Fed. Probation, 80:38, 2016.

[21] I. Goodfellow, Y. Bengio, and A. Courville. Deep learning. MIT press, 2016.

[22] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

[23] P. Hall and N. Gill. Debugging the black-box compas risk assessment instrument to diagnose and remediate bias. ICML Open-Review 2017, 2017.

[24] M. Hardt, E. Price, N. Srebro, et al. Equality of opportunity in supervised learning. In Advances in neural information processing systems, pages 3315–3323, 2016.

[25] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. Neural networks, 2(5):359–366, 1989.

[26] C.-L. Huang, M.-C. Chen, and C.-J. Wang. Credit scoring with a data mining approach based on support vector machines. Expert systems with applications, 33(4):847–856, 2007.

[27] A. E. Khandani, A. J. Kim, and A. W. Lo. Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34(11):2767–2787, 2010.

[28] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[29] J. Kleinberg and S. Mullainathan. Simplicity creates inequity: Implications for fairness, stereotypes, and interpretability. arXiv preprint arXiv:1809.04578, 2018.

[30] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. How we analyzed the compas recidivism algorithm. ProPublica (5 2016), 9, 2016.

[31] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt. Delayed impact of fair machine learning. arXiv preprint arXiv:1803.04383, 2018.

[32] K. Lum and W. Isaac. To predict and serve? Significance, 13(5):14–19, 2016.

[33] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.

[34] C. Munoz, M. Smith, and D. Patil. Big data: A report on algorithmic systems, opportunity, and civil rights. Executive Office of the President, 2016.

[35] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against machine learning. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, pages 506–519. ACM, 2017.

[36] R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V. McConnell, G. S. Corrado, L. Peng, and D. R. Webster. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nature Biomedical Engineering, 2(3):158–164, 2018.

[37] ProPublica. COMPAS Analysis. GitHub, 2016. https://github.com/propublica/compas-analysis.

[38] S. K. Sarkar, K. Oshiba, D. Giebisch, and Y. Singer. Robust classification of financial risk. arXiv preprint arXiv:1811.11079, 2018.

[39] J. Suk Gersen, M. Haley, D. Wegner, M. Lin, R. Millett, M. Brenner, S. Suproteem, and A. Merchant. Opening the blackbox: Constitutional dimensions of predictive algorithms in criminal justice. Working paper, 2019.

[40] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.

[41] C. Wadsworth, F. Vera, and C. Piech. Achieving fairness through adversarial learning: an application to recidivism prediction. arXiv preprint arXiv:1807.00199, 2018.

[42] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. arXiv preprint arXiv:1801.07593, 2018.