
DATAJUSTICE
ERC-funded project 'Data Justice: Understanding datafication in relation to social justice' (DATAJUSTICE) starting grant (2018-2023).

WORKING PAPER

# How to (partially) evaluate automated decision systems

Javier Sánchez-Monedero and Lina Dencik
Sanchez-MonederoJ,DencikL at cardiff.ac.uk
Cardiff University

December 6, 2018

datajusticeproject.net

# Contents

# 1 Introduction

Depicting the social impact of automated decision systems requires multiple interdisciplinary entry-points. In this paper we focus on the actual data and algorithms that produce specific outputs for the purposes of decision-making. The aim of this report is to outline the range of prominent methods that are used for auditing algorithms in data-driven systems and to also consider some of their limitations.

A number of early academic studies on discrimination in machine learning and statistical modelling were published in 2010-2011 (Ruggieri et al., 2010; Luong et al., 2011; Pope and Sydnor, 2011), but the topic received prominent attention in public debate with the investigation carried out by Propublica on the COMPAS system used in parts of the criminal justice system in the United States (Angwin and Larson, 2016). Whilst the original publication has been the focus of most discussions pertaining to data-driven discrimination, the re-evaluations of ProPublica's work have received much less attention. As a consequence, there has been an all too hasty engagement with the design of solutions to the problem without proper scrutiny of the study or the multiple contextual factors at play. Due to the different definitions of fairness, the lack of access to public data, documentation and code, the results of a number of studies re-evaluating this case raise different or even contradictory conclusions with respect to the COMPAS system. In this report, we will pay particular attention to this case as a way to illustrate methods of evaluating and auditing algorithmic decision-making.

# 2 Brief introduction to machine learning and some cautions

Machine learning (ML), also known as statistical learning, is the field of study that gives computers the ability to learn from data in order to perform a task without being explicitly programmed. That is, to build/learn/fit a model from data. Tom M. Mitchell defined a ML algorithm as '*A computer program [that] is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E*' (Mitchell, 1997). In general, this definition will be helpful to organise the auditing of a learning system: What is the **actual task** T that is being learned? How is the **experience** E represented as quantitative data? How is the **performance** P measured to control the learning process and to evaluate the final model?

Imagine we build an *intelligent* system to help people decide whether or not to buy a house for a particular price. In the following example, the goal is to predict the average price of an apartment (T), using a dataset of percentage of AirBnB apartments in a city. So a model can learn to predict the average price of apartments in a city using the % of AirBnB in the city, this is our dataset of experience (E). Learning or model fitting consist on an iterative process which changes a model to find the minimum error in the prediction (P) using the training dataset (see Figure 1).

This simple example gives us some hints about issues related to the design of machine learning based solutions, and consequently how to perform a study based on public documentation. What is the selection of variables? How is each variable produced? Is the task solved by the machine learning algorithm the same as the task the system is aiming to solve? What are the limitations of the model we are using (in the example we assume a linear relationship between variables)? What metric are we considering to guide and evaluate the learning process? What is the learned model revealing about the task?

# 3 The COMPAS case

## 3.1 The ProPublica analysis and initial findings in the COMPAS case

COMPAS (which stands for Correctional Offender Management Profiling for Alternative Sanctions) is an automated tool developed by Northpointe, Inc. to predict a score to determine whether

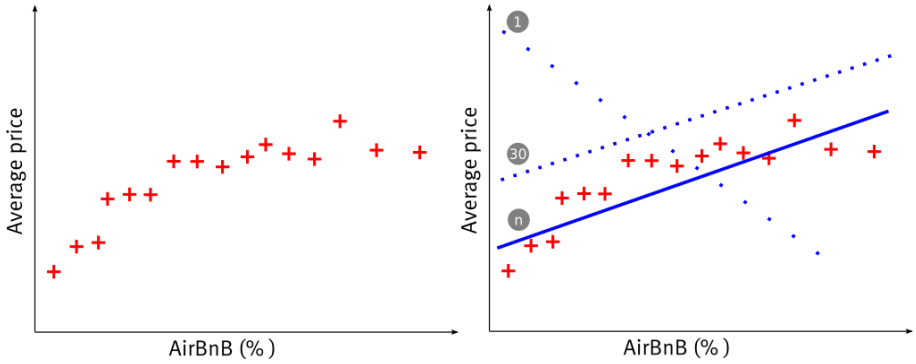



Figure 1: Example of a regression task and model fitting process for linear regression.

to release or detain a defendant before his or her trial. The tool is used across the US and based on a set of features it automatically determines a score ("Risk of Recidivism," "Risk of Violence" and "Risk of Failure to Appear") for each pre-trial defendant.

The COMPAS tool uses a set of variables describing a person such as criminal history, charge degree, gender, race or age (independent variables) to produce a score (dependent variable). Rather than being directly programmed by a human, this is done by training several statistical models with historical records of criminals (that is, independent variables and dependent variable patterns) to build a predictive model for the score. By doing so, the general assumption is that the statistical model will not discriminate by any sensitive attribute such as race or gender.

However, the tool has been at the centre of a public and scientific debate since the well-known research work of ProPublica claiming racial bias of the tool, which, according to their analysis, is more prone to penalise black defendants with respect to white ones in several ways (Angwin and Larson, 2016). Rather than being explicitly programmed to discriminate against black individuals, the statistical models learned to incorporate the historical bias present in the data records. To claim this, the ProPublica team replicated, as far as they could, the Northpointe Inc. tool according to the public documentation of the software and public records of criminality.

The main finding can be summarised as the over-estimation of risk predicted for black defendants which directly impacts on their chances of bail. Analogously, white defendants were found to have more chances of being evaluated as low risk of recidivism.

### 3.2 Controversy around the ProPublica analysis and complementary studies

Northpointe's Research Department published a technical report in which they '*strongly reject the conclusion that the COMPAS risk scales are racially biased against blacks*' (Dieterich et al., 2016). Then, ProPublica published a response to this report (Larson and Angwin, 2016b). Northpointe pointed out a series of several statistical and technical errors such as mis-specified regression models, wrongly defined classification terms and measures of discrimination, and the incorrect interpretation and use of model errors.

The first critique of ProPublica from Northpointe is that they used different samples to conduct the three different analyses. In addition, ProPublica does not include descriptive statistics about the sample. Therefore, it is not possible to properly value information such as different scores ratios and different error ratios for black individuals. Related to this concern, they say the study falls into the base rate fallacy[1] by not considering the frequency of the events in the population (formally prior probability) to correct the statistical analysis. In other words, if there are more high risk black defendant profiles in the database, the performance metrics will be affected by this fact, so the model performance evaluation should be corrected using the Bayes theorem to estimate the posterior probability as a more robust performance estimator. However, ProPublica

[1] https://en.wikipedia.org/wiki/Base_rate_fallacy

partially avoid this issue since they present the confusion matrices together with a variety of performance metrics to better assess the overall performance. These metrics are already scaled by the frequency of the recidivism events for each group. Finally, Northpointe claims that their system uses much more variables (137 features which do not include any race variables) than the ProPublica simulation (12 features). However, it has been noted that even with less variables, the ProPublica model performance is aligned with the Northpointe model performance, questioning the necessity of remainder features. Also, as we go on to discuss below, a later study on COMPAS by Dressel and Farid (2018) demonstrated that simple models with two variables can achieve the same performance.

It is worth pointing to the argument presented by Corbett-Davies et al. (2016) illustrating the key issue that the controversy raises in the definition of fairness. Northpointe's definition of fairness is to predict a similar proportion of defendants that reoffend within each risk category ("low", "medium", "high"). On the other hand, ProPublica claims that it is unfair that '*black defendants who don't reoffend are predicted to be riskier than white defendants who don't reoffend*'. According to the authors, the point is that both notions of fairness are mathematically guaranteed within the same scenario.

Apart from the Northpointe report, other independent analyses of recidivism systems have been performed. Some of them only relied on statistical analysis of the case, but fortunately others went beyond the technical questions. The rest of this section summarises some of these studies and reveals the necessity of more wide and interdisciplinary analyses of data-driven systems.

Skeem and Lowenkamp (2016) specifically studied the predictive bias and disparate impact related to race in the PostConviction Risk Assessment (PCRA). The PCRA was developed for federal offenders in the US, who differ from state-level offender profiles, but still in the field of recidivism evaluation. They identified the same probability of recidivism across groups whilst black individuals obtain higher average PCRA. The authors attribute this disparate impact to the difference in the criminal history but they do not consider the criminal history as a proxy variable to race, but a variable that instead mediates[2] the relationship between race and future arrest.

Zhang and Neill (2016) presented a method for subset scanning to detect statistical significant bias in binary classifiers. The method is able to describe the characteristic of the discriminated group by testing for bias or poor fitting regions in the mathematical model. They raise new conclusions of the COMPAS case: rather than identifying bias among clearly defined racial groups, the method found a multi-dimensional subgroup: '*females who initially committed misdemeanors (rather than felonies), for half of the COMPAS risk groups, have their recidivism risk significantly overestimated.*'.

Gong (2016) re-analysed the ProPublica study and concluded that '*The COMPAS-to-recidivism lines for black and white offenders are closely aligned over the full range of scores.*'. However, in his text, he goes further than the technical analysis, with a powerful conclusion: '*In a way, that's fortunate, because it creates an opportunity to look at how powerful algorithms can be deeply unfair, even when they're statistically unbiased.*'. For instance, Gong set the focus on the definition of the machine learning task; what the systems are really predicting. In the case of COMPAS, the system is not predicting future crimes, but *arrest* for future crimes.

More recently, Dressel and Farid (2018) studied the fairness performance of COMPAS excluding racial variables (a point of view surprisingly missing in the previous studies). However, apart from confirming the behaviour revealed by ProPublica (different false positive and false negative rates across groups), they also evaluated the performance and fairness of non-expert humans vs. computers in the COMPAS risk assessment by providing the same variables for both forms of decision-making. Based on this experiment, they conclude: '*We show... that the widely used commercial risk assessment software COMPAS is no more accurate or fair than predictions made by people with little or no criminal justice expertise. In addition, despite COMPAS's collection of 137 features, the same accuracy can be achieved with a simple linear predictor with only two features.*'. The conclusions of this study raise several interesting questions, particularly as debates on discrimination in algorith-

[2]For the definition of mediation in statistics see https://en.wikipedia.org/wiki/Mediation_(statistics).

**Table 1. Human versus COMPAS algorithmic predictions from 1000 defendants.** Overall accuracy is specified as percent correct, AUC-ROC, and criterion sensitivity ($d'$) and bias ($\beta$). See also Fig. 1.

| | (A) Human (no race) | (B) Human (race) | (C) COMPAS |
|---|---|---|---|
| Accuracy (overall) | 67.0% | 66.5% | 65.2% |
| AUC-ROC (overall) | 0.71 | 0.71 | 0.70 |
| $d'/\beta$ (overall) | 0.86/1.02 | 0.83/1.03 | 0.77/1.08 |
| Accuracy (black) | 68.2% | 66.2% | 64.9% |
| Accuracy (white) | 67.6% | 67.6% | 65.7% |
| False positive (black) | 37.1% | 40.0% | 40.4% |
| False positive (white) | 27.2% | 26.2% | 25.4% |
| False negative (black) | 29.2% | 30.1% | 30.9% |
| False negative (white) | 40.3% | 42.1% | 47.9% |

Figure 2: Comparison of human versus algorithmic prediction. Source (Dressel and Farid, 2018).

mic decision-making is often framed around this (problematic) binary of humans vs. computers. On the one hand, they reveal that discriminatory bias is prevalent across both forms of decision-making (even when excluding the race information) while at the same time they question the unnecessary complexity of some data-driven systems. See Figure 2 for details.

# 4 Methods for auditing automated decision systems

The rest of this working paper presents several methods used to study automated decision systems. We first present the initial problem of how to perform a study in the absence of the actual data and code of the system. Then, we present some studies analysing the problem at the data level. We continue summarising several ways of detecting bias analysing both the algorithm's output and internal behaviour. The section ends with some public resources to perform analyses of ML models.

## 4.1 Replicate behaviour using other systems and data

Ideally, auditing a product would include having access to data, source code and internal documentation of the whole system (including third party components), as well as interviewing the design and development team. However, this is far away from the common settings of all the case studies we present in this working paper.

In the absence of such ideal situation, to analyse potential harms of some systems, many researchers are replicating the tools according to public documentation. Likewise, in absence of the original dataset and/or ground truth data, similar data is gathered to try to replicate the decision algorithm as a way to analyse its behaviour.

For example, this is what ProPublica did for the COMPAS tool (Larson and Angwin, 2016a). They collected COMPAS scores and criminal records from the Broward Country. The raw data of COMPAS scores and criminal records included more than 50 variables describing each case, and it was preprocessed to create the final 12 features of the statistical model. For instance, the

COMPAS score ranges from 1 to 10 and scores 1 to 4 were labeled as "Low"; 5 to 7 were labeled "Medium" and 8 to 10 were labelled "High". For the purposes of binary analysis, the "Low" label was considered the positive class and labels "Medium" and "High" were the negative class. The age was not directly considered, but instead two binary variables were introduced to represent if a person is less than 25 or more than 45.

However, some case studies lack "ground truth" datasets. For instance, in the case of predictive policing to identify hot-spots of crimes, there is not data about real crimes across areas, since the only data available are the police records. Therefore, evaluating bias in this data (the police activity) is not straightforward. This is the case of the study of Lum and Isaac (2016) of the predictive policing tool PredPol in Oakland (California). The authors selected the area because PredPol was used there and there was publicly available open data[3]. To investigate the effect of police recorded data on predictive policing models, they applied the algorithm of the PredPol software that was previously described in a journal paper. In the absence of ground truth data, they compared the map of drug arrests made by Oakland police with a drug users map generated with a synthetic population, matching the demographic features of the population in Oakland's crime records (see Figure 3a). Figure 3b (top) shows the number of days PredPol was targeting different areas. The comparison of drug user map in Figure 3a and crime prediction map in Figure 3b revealed that the tool was targeting mainly two areas with largely non-white and low-income populations while the drug crimes were much more evenly distributed across the city. Also, that neighbourhoods with a higher proportion of ethnic minorities experienced about 200 times more drug-related arrests than white and low-income populations (see Figure 3b).

## 4.2 Descriptive statistics and data visualization

Applying basic qualitative analysis, descriptive statistics and data visualization are the common first steps in systems auditing. At the data collection and evaluation stage, *bias* will typically be referred to in the statistical sense: there is a difference between the observation and the reality, which does not imply judgement.

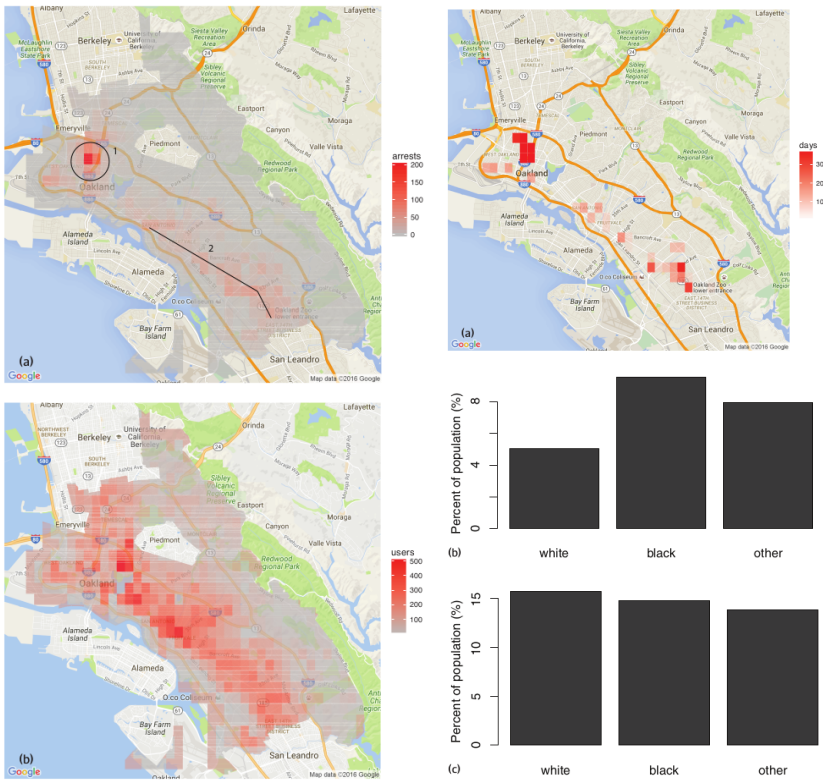
For instance, Price and Ball (2014) analysed reports of violent events and mortality in Syria and Iraq to identify bias in human rights data collection. Their study focuses on a kind of selection bias called *event size bias*: '*Event size bias is the variation in the probability that a given event is reported, related to the size of the event: big events are likely to be known, small events are less likely to be known. In studies of conflict violence, this kind of bias arises when events that involve only one victim are less likely to be documented than events that involve larger groups of victims.*' This can also be related to the public nature of the attack, for instance a market bombing attack.

To look for event size bias, the authors gathered victims data reports from four different Syrian sources, and represented the number of victims in a temporal line in May 2013 (see Figure 4). They found a correlation between the type of event and the likelihood of it being reported, which can lead to misleading conclusions if the bias is not adjusted.

In the ProPublica analysis on machine bias, histograms are used as a first step to analyse the distribution of the frequency of risk of recidivism across whites and black defendants. This simple analysis can reveal issues with the data collection, (apart from highlighting discrimination in the criminal system in the US). From a probabilistic point of view the skewed distribution of Figure 5 means that the prior probability of high risk will be higher for black defendants. As a consequence, any machine learning algorithm having access to the race variable, or proxy variables[4] for that variable, are very likely to learn to assign higher risk to black individuals. Considering the learning algorithms, this happens since the race variable will effectively help the learned model to minimize the global classification error for this specific (biased) dataset. Finally, histogram analysis can also summarise demographics of data features, and then have to be considered in posterior analyses of the performance of the model with respect to groups.

[3] https://hrdag.org/2016/11/04/faqs-predpol/

[4] https://en.wikipedia.org/wiki/Proxy_(statistics)




(a) Number of drug arrests (top) compared to the estimation of drug users (bottom).

(b) Number of days with targeted policing for drug crimes in areas flagged by PredPol (top). Race distribution of targeted policing.

Figure 3: Contrast of sources of information in absence of ground truth and analysis of the behaviour of PredPol. Source (Lum and Isaac, 2016).

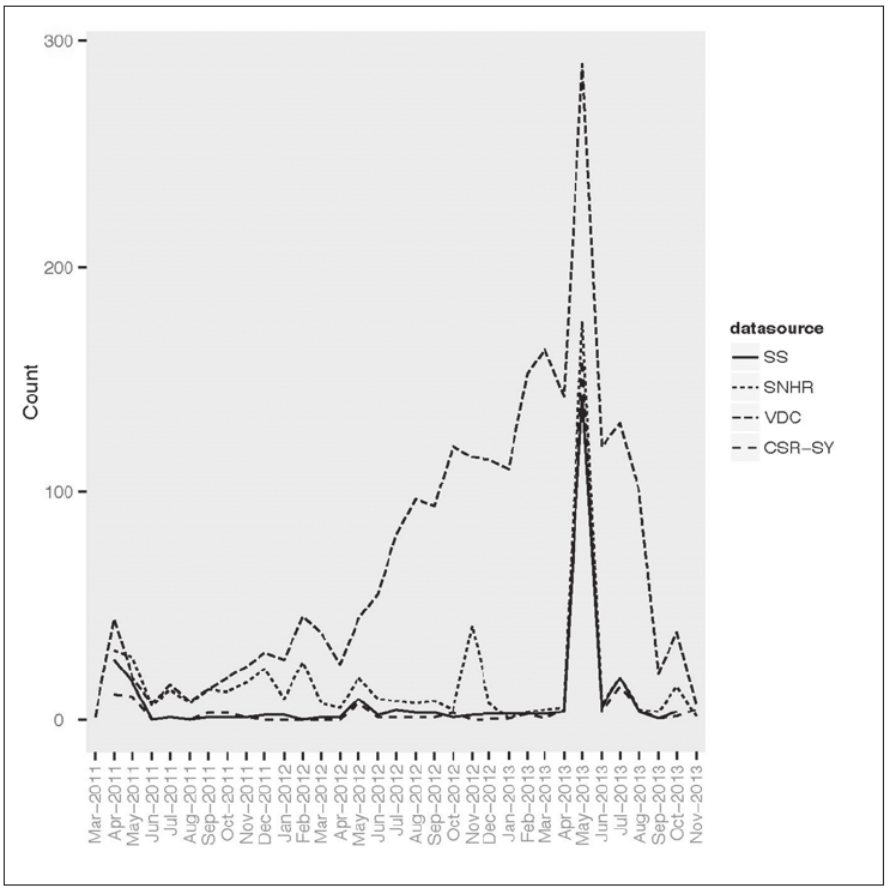



Figure 4: Example of event size bias. Large events are more likely to be reported by many organizations. Source (Price and Ball, 2014).

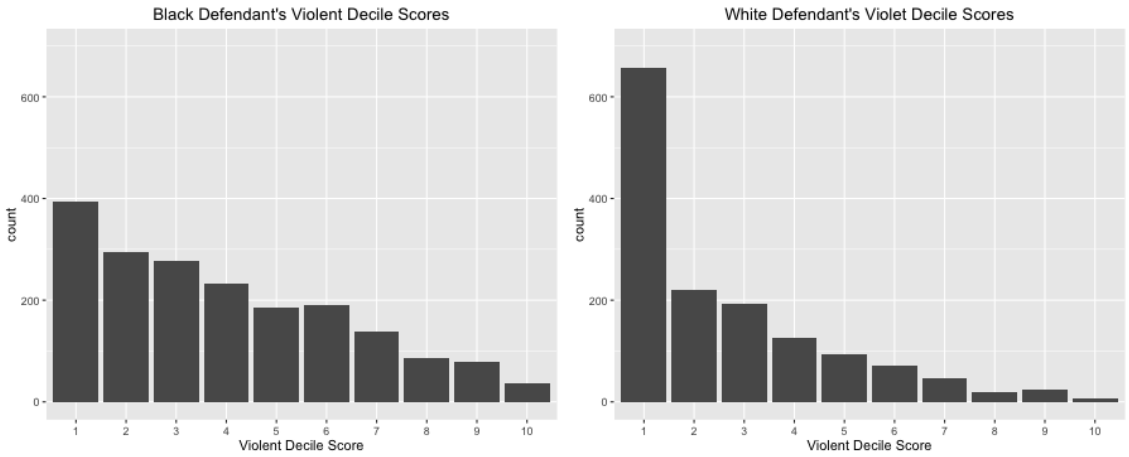



Figure 5: Decile scores for black (left) and white (right) defendants in the COMPAS database. Source (Larson and Angwin, 2016a).

### 4.3 Algorithm performance analysis and statistical fairness

As mentioned, the machine learning process is guided by loss functions and it is evaluated using performance metrics such as the global accuracy, the ROC (receiver operating characteristic) curve analysis or the sensitivity/specificity statistical measures. The selection of the specific loss function can significantly impact on the resulting model and the error metrics generally affect the model selection and hyper-parameter tuning. Altogether this conditions the generalization behaviour of a model. Moreover, the performance metric selection will highlight different aspects of the model whilst at the same time can be strategically used to hide algorithmic bias.

Assessing discrimination performance from different points of view[5] is general good practice in statistics, and in areas such as medicine it is a must to properly illustrate a classifier behaviour. For instance, ProPublicas's claim of unfairness in the case of COMPAS is based on highest false positive rate (FPR) for black defendants, meaning people that did not reoffend were miss-labelled as high risk profiled people. It turned out that the FPR was 44.85% for black individuals whilst 23.45% for white individuals (see Figure 6). On the other hand, Northpointe's notion of fairness is to achieve the same predictive positive value (PPV) for white and black defendants.

Many of the technical studies and fairness proposals in the literature are based on identifying and mitigating algorithms' unfairness based on the notion of *statistical parity*, also called group parity, with respect to different performance metrics. These concepts have been technically defined in the context of classification or content retrieval (Žliobaitė, 2017).

Performance comparison of subgroups can also be evaluated visually with the receiver operating characteristic (ROC) curve representation for two or several groups (see Figure 7).

The above analysis is valid to compare a binary classification task (for instance 'jail' vs 'release' prediction). However, to compare performance among multiple labels requires other approaches. This situation is quite common in computer vision problems of image labelling: provided with an image, try to label the objects and actions in a picture ("car", "woman', "man", "tree", "dog", "cook"...). To identify bias in such scenarios, Zhao et al. (2017) propose a bias score of a given output ("car", "kitchen", "shopping"...) with respect to a demographic variable such as gender. Also, in this work they measure the bias amplification by comparing the bias score in the training dataset with the bias score of the model predictions in the generalization set (assuming identical distribution in the train and test sets). Zhao et al. tested the bias visualization in two problems of role labelling (vSRL) and multilabel image classification (MLC). Figure 8 presents the visualisation proposal by Zhao et al. (2017).

Finally, Hardt et al. (2016) propose several strategies to achieve equality of opportunity in candidate selection in machine learning. Although their article is centred on a case of loan granting or denying, the study can be generalised to other contexts. In the article, they present several

[5] https://en.wikipedia.org/wiki/Confusion_matrix

| All Defendants | | | Black defendants | | | White defendants | | |
|---|---|---|---|---|---|---|---|---|
| | Low | High | | Low | High | | Low | High |
| Survived | 4121 | 1597 | Survived | 1692 | 1043 | Survived | 1679 | 380 |
| Recidivated | 347 | 389 | Recidivated | 170 | 273 | Recidivated | 129 | 77 |
| FP rate: 27.93 | | | FP rate: 38.14 | | | FP rate: 18.46 | | |
| FN rate: 47.15 | | | FN rate: 38.37 | | | FN rate: 62.62 | | |
| PPV: 0.20 | | | PPV: 0.21 | | | PPV: 0.17 | | |
| NPV: 0.92 | | | NPV: 0.91 | | | NPV: 0.93 | | |
| LR+: 1.89 | | | LR+: 1.62 | | | LR+: 2.03 | | |
| LR-: 0.65 | | | LR-: 0.62 | | | LR-: 0.77 | | |

Figure 6: Performance evaluation considering groups of COMPAS predictive model reproduced by ProPublica. Source (Larson and Angwin, 2016a).

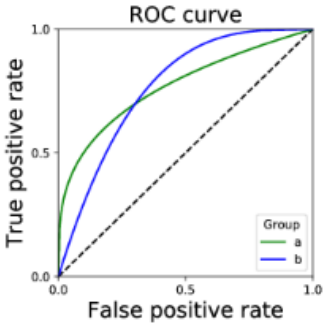
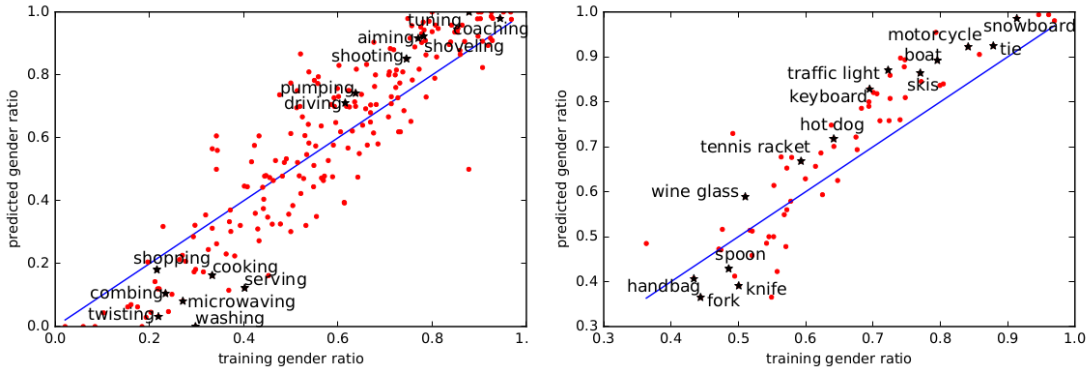



Figure 7: ROC curve comparison of the performance of a binary classification model with respet to two groups. Source Barocas and Hardt (2017).




(a) Bias analysis on imSitu vSRL

(b) Bias analysis on MS-COCO MLC

Figure 2: Gender bias analysis of imSitu vSRL and MS-COCO MLC. (a) gender bias of verbs toward `man` in the training set versus bias on a predicted development set. (b) gender bias of nouns toward `man` in the training set versus bias on the predicted development set. Values near zero indicate bias toward `woman` while values near 0.5 indicate unbiased variables. Across both dataset, there is significant bias toward males, and significant bias amplification after training on biased training data.

Figure 8: Visual comparison of model discrimination in multi-label problems. Proposal by (Zhao et al., 2017).

threshold techniques to satisfy different loan strategies (max profit for the bank, ignore groups performance, demographic parity and equal opportunity). The behaviour of the model can be explored in an interactive website[6]. One of the interesting contributions of this work is the clear connection of different high-level policies or strategies with the technical solutions. For instance, the high-level goal of maximizing total profit will produce decision thresholds that will minimize credit defaults, no matter if it is at the cost of penalising a group. On the other hand, each group threshold has to be different to achieve a high-level criterion of equal opportunity, meaning equal performance for all groups for right credit loans among people who would pay back (see Figure 9).

## 4.4 Model interpretability

Although there is a general belief that all machine learning and AI models are black-box models, many data-driven systems are still relying on classic statistical models such as logistic regression

[6] http://research.google.com/bigpicture/attacking-discrimination-in-ml/

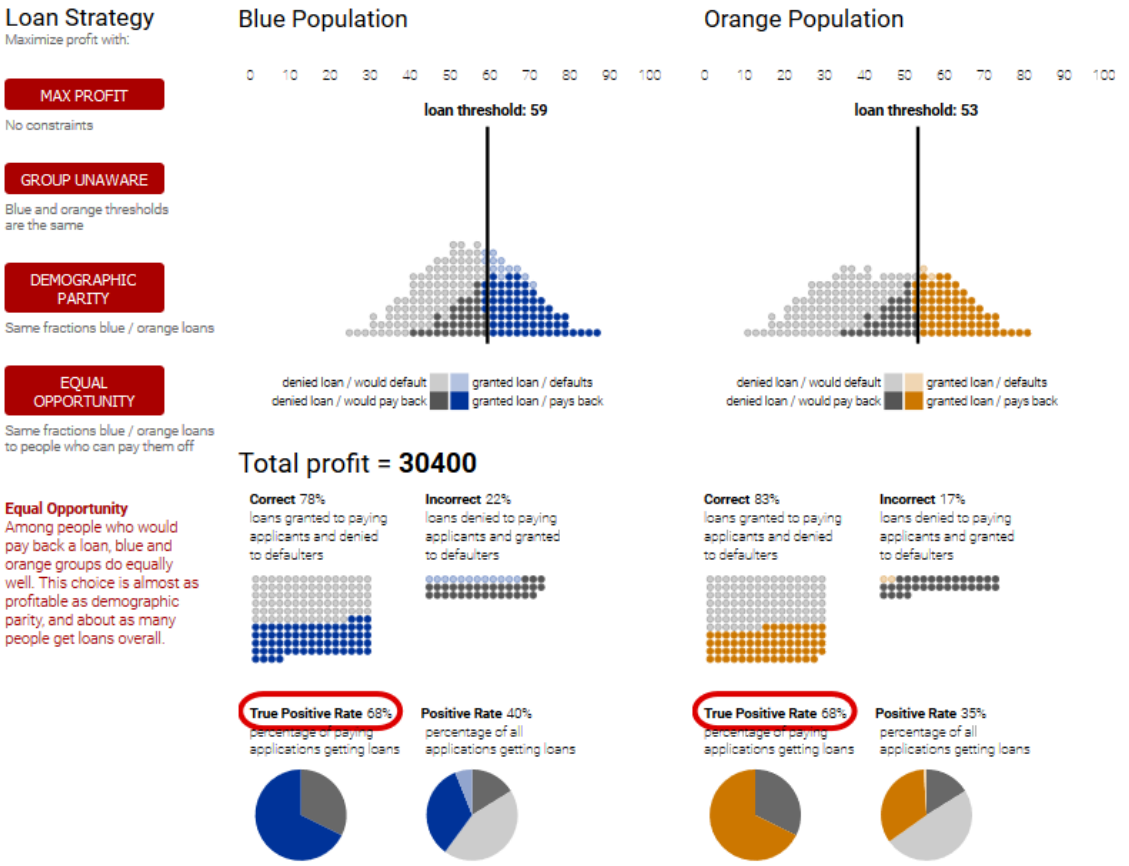



Figure 9: Implementation of different threshold calibration to achieve high-level criterion. Scenario of equal opportunity loan strategy. Created with the loan threshold simulator associated to the work of (Hardt et al., 2016).

because of the model interpretability and performance for large datasets among other reasons. Also, other methods such as decision trees[7] or Bayesian networks[8] are suitable for internal analysis.

In general terms, generalised linear model fitting (also known as model training or learning) consist of adjusting weights for each variable to better fit into the training data whilst trying to minimise a loss function that represents how far the model predictions are from the actual labels. Therefore, models such as the logistic regression are widely used in many fields not only to build classifiers, but also to analyse the contribution of independent variables to the probability of a success (positive class). For instance, the table in Figure 10 visualises the weight of each variable of a logistic regression model trained with the COMPAS dataset (here the success, or positive class, is a "medium and high risk" of recidivism). A positive weight of the value means that it contributes directly to attributing a high risk score to the person whilst negative values decreases the probability of high risk. For instance, from the above analysis we can assert that being under 25 is more relevant than the number of priors, and the same applies for the variable 'Black' with respect to the number of priors. This means, that a young black person with no priors has more likelihood of being classified as a risky person than an old white man with several priors. ProPublica presented some concrete profiles of inflated risk scores by this statistical model in their press article (Angwin and Larson, 2016).

On the other hand, some works are focused on auditing black-box models. The work of Mol-

[7] http://www.r2d3.us/visual-intro-to-machine-learning-part-1/
[8] https://en.wikipedia.org/wiki/Bayesian_network

**Risk of General Recidivism Logistic Model**

| | *Dependent variable:* |
|---|---|
| | Score (Low vs Medium and High) |
| Female | 0.221*** (0.080) |
| Age: Greater than 45 | -1.356*** (0.099) |
| Age: Less than 25 | 1.308*** (0.076) |
| Black | 0.477*** (0.069) |
| Asian | -0.254 (0.478) |
| Hispanic | -0.428*** (0.128) |
| Native American | 1.394* (0.766) |
| Other | -0.826*** (0.162) |
| Number of Priors | 0.269*** (0.011) |
| Misdemeanor | -0.311*** (0.067) |
| Two year Recidivism | 0.686*** (0.064) |
| Constant | -1.526*** (0.079) |
| Observations | 6,172 |
| Akaike Inf. Crit. | 6,192.402 |

*Note: *p<0.1; **p<0.05; ***p<0.01*

Figure 10: Variable scores of the COMPAS logistic regression model of the ProPublica study. Source (Larson and Angwin, 2016a).

nar (2018) provides an extensive guide of techniques for explaining black-box models. For instance, Adebayo and Kagal (2016) propose a model-agnostic method to analyse the relevance of each variable in the classification prediction. This makes it possible to perform an analysis similar to the logistic regression analysis we have seen. Beyond model internals, other methods can provide explanations by looking for similar or synthetic datapoints that can be used as class prototypes as well as counterfactual examples. Interpretability can also be understood as the capability o1f explaining individual decisions. For instance, local surrogate models (LIME) by (Ribeiro et al., 2016) can be used to approximate explanations of individual predictions. LIME can be applied to tabular, text and image data (See Figure 11).

## 4.5 Public tools to evaluate machine learning models and achieve statistical fairness

In recent years, several tools have been released to perform machine learning auditing, bias discovery of fairness-aware machine learning methods. The following open source tools are available for model auditing[9]:

- Fairness Measures (Meike et al., 2017) implements the fairness evaluation metrics defined by (Žliobaitė, 2017) and provides a set of datasets to evaluate bias issues in learning algorithms. The project focuses on classification and ranking algorithms[10].
- The Algorithmic Fairness group at Haverford College[11] maintains a public repository on

[9]Note we do not include privative and closed solutions. Also, we do not enumerate all the open source projects related to fairness evaluation. Other excellent repositories are Audit-AI, Fairness Comparison, Fairness, FairTest, Themis$^{TM}$, and Themis-ML.

[10]http://fairness-measures.org

[11]http://fairness.haverford.edu/

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 11: Example of explanation of image classification generated by LIME. Source (Ribeiro et al., 2016).

algorithmic fairness[12] including fairness benchmarking (Friedler et al., 2019), black box auditing (Feldman et al., 2015; Adler et al., 2018) as well as code related to fairness-aware method proposals (Ensign et al., 2017).

- Aequitas[13] is an open source bias audit toolkit for machine learning developers, analysts and policymakers. The toolkit has been developed for classification tasks by the Center for Data Science and Public Policy[14] at University of Chicago. The tool generates reports that can be understood by a non-technical audience[15].
- LIME[16] (local interpretable model-agnostic explanations) is a project to provide explanations on individual predictions of text, table-based and image-based classifiers.
- FairML[17] is a python toolbox auditing the machine learning models to identify the significance of each variable to the classification model.
- What-If Tool[18] by Google is a code-free tool to analyse decisions of machine learning models built on TensorFlow[19]. The tool's features are prediction evaluation, synthetic data evaluation, exploration of single feature effects, comparison of counterfactual examples (similar data points with different labels), arrangement of examples by similarity and testing algorithmic fairness constraints.
- AI Fairness 360[20] is an open source software toolkit by IBM. The toolbox is especially interesting since it can be integrated in the machine learning pipeline to regularly check for unwanted biases and to mitigate any biases that are discovered[21].
- Fairness in Classification[22] repository provides a logistic regression implementation in python for the fair classification mechanisms introduced in several papers (Zafar et al., 2015, 2017a,b). The repository includes several fairness evaluation metrics.

[12] https://github.com/algofairness
[13] https://dsapp.uchicago.edu/projects/aequitas/ and https://github.com/dssg/aequitas
[14] https://dsapp.uchicago.edu/
[15] Report example at http://aequitas.dssg.io/example
[16] https://github.com/marcotcr/lime
[17] https://github.com/adebayoj/fairml
[18] https://pair-code.github.io/what-if-tool/
[19] https://www.tensorflow.org/
[20] http://aif360.mybluemix.net/
[21] https://developer.ibm.com/code/open/projects/ai-fairness-360/
[22] https://github.com/mbilalzafar/fair-classification/

# 5 Conclusions

This working paper does not cover some topics and studies. For instance, some system auditing can be performed by simply accessing the system interface. ProPublica did this to test whether Facebook allow racial filters in targeted advertisement (Angwin and Tobin, 2017). We did not cover patents study as a means to obtain insights of some data systems. In the field of machine learning, we did not cover the area of ranking or recommendation engines or discrimination discovery methods, for instance the ones based on rules discovery (Ruggieri et al., 2010).

The COMPAS case is the best documented and studied case when it comes to the question of discrimination and fairness in data and machine learning. However, none of the teams auditing the software have access to the actual data or the code or mathematical models of the system. Therefore, their conclusions are based on reproducibility of experiments with a scarcity of resources. Thus, the actual system assessing risks of real persons remains unknown. This remains a major challenge in the auditing of algorithms.

Biases of machine learning models have often been hidden under global performance metrics. Many academic works have only relied on accuracy (global performance), sensitivity and specificity analysis (and ROC analysis) considering the class labels. However, fairness across social groups in machine learning is a recent emerging topic that has received particular attention after ProPublica's COMPAS study. For instance, numerous open source toolkits have been developed to implement performance group aware metrics. Whilst it is difficult (or even impossible) to produce a (general) statistical definition of fairness, the media and policymakers have been seduced by global performance results of data-driven systems ('it's 95% accurate!') to justify the data-driven systems adoption. Yet cases such as the 'staggeringly inaccurate' facial recognition tools deployed by police in Wales[23] become less surprising if concepts such as the base ratio fallacy or the Bonferroni's Principle[24] are considered to contextualise the evaluation.

Attention should be paid to the data acquisition and features design as well as data labelling. It is very intuitive to think that a criminal historic record with persons labelled (dependent variable) by humans can be biased with respect to marginalized groups. However, features design can hide biases under more subtle representations. For instance, Skeem and Lowenkamp (2016) call attention to the coding of historical records that mediates (influences) the relation of the independent variable with the dependent variable in a different negative manner for black individuals. This is reinforced by the study of Dressel and Farid (2018) in which the false positive ratios are still different for black and white individuals even when the race information is hidden. This behaviour was found both for computers and humans performing a risk evaluation task based only on criminal records. This evidence is aligned with the analysis of Harcourt (2010), which argues that risk is a proxy for race. Harcourt's argument is focused on the fact that risk is related to prior criminal history, and prior criminal history has become a proxy for race in the US. Finally, the tutorial on Fair ML by Barocas and Hardt (2017) emphasizes the relevance of measurements, '*the #1 neglected topic in statistics*' (Gelman, 2015), and concludes that'*Social questions start with measurement*'.

Transparency and interpretability of algorithms and models are also key points. The logistic regression analysis is the predominant accepted method. However, some cautions should be considered. Among others, the contribution to the probability of the success assumes a linear relationship between independent and dependent variables. Also, this type of model does not capture variable interactions. In addition, it is worthwhile to point out here that model analysis is performed in a model built using a training set with a specific loss function. The relationship of this loss metric, that guides the learning process, to the problem that the model is assumed to solve has to be carefully examined. From this, we can conclude that the interpretability process is tied to the specific problem and needs human expertise to carefully extrapolate conclusions. We have seen some works on model interpretability regarding variable weight, performance evaluation, counterfactual examples, relevant regions of the input space, etc. but the topic of interpretable machine learning is nevertheless a field of study in its own right (Molnar, 2018). Whilst inter-

[23] https://www.bbc.co.uk/news/technology-44089161

[24] https://rationalwiki.org/wiki/Bonferroni%27s_principle

pretability of models is often claimed, prior definition, scope and limitation is generally missed. We refer to the paper of Lipton (2016) for further discussions on this topic.

Even if we bound the discussion into the merely technical aspects, the general data context of the system remains under-analysed in the literature and public debate and would lead to questions such as: what are the input variables and how are they produced; what is the output of the model (i.e. what it is actually predicting) and what is being optimised by the learning algorithm to build the classification model? These pertinent questions heavily condition the model behaviour analysis and conclusions. For instance, in the COMPAS problem, rather than predicting crimes, the model is predicting whether a person will reoffend and will be detained by the police (both simultaneous events). The learning algorithm will learn to evaluate the risk that 'a person will reoffend and be caught by the police', hence the interpretation of performance analysis and variable weight changes and therefore the conclusions. We already mentioned the issue of criminal records and risk as race proxies. Would we have different impact with different representations of information? Can we build fair systems based on the current feature selection and measures? Can we achieve social progress in some domains by building models on historical data? This wider discussion is generally missed in public and academic debate and points precisely to the questions that can reveal the role of each data-driven system in society.

## References


Julius Adebayo and Lalana Kagal. Iterative Orthogonal Feature Projection for Diagnosing Bias in Black-Box Models. *arXiv:1611.04967 [cs, stat]*, November 2016. URL http://arxiv.org/abs/1611.04967. arXiv: 1611.04967.

Philip Adler, Casey Falk, Sorelle A. Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing Black-box Models for Indirect Influence. *Knowl. Inf. Syst.*, 54(1):95–122, January 2018. ISSN 0219-1377. doi: 10.1007/s10115-017-1116-3. URL https://doi.org/10.1007/s10115-017-1116-3.

Julia Angwin and Jeff Larson. Machine Bias. *ProPublica*, May 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Julia Angwin and Ariana Tobin. Facebook (Still) Letting Housing Advertisers Exclude. . . , November 2017. URL https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin.

Solon Barocas and Moritz Hardt. Fairness in Machine Learning. NIPS 2017 Tutorial, 2017. URL http://fairml.how/.

Sam Corbett-Davies, Emma Pierson, Avi Feller, and Sharad Goel. A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear., October 2016. URL https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/.

William Dieterich, Christina Mendoza, and Tim Brennan. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Performance of the COMPAS Risk Scales in Broward County. Technical report, Northpointe Inc., July 2016. URL http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580, January 2018. ISSN 2375-2548. doi: 10.1126/sciadv.aao5580. URL http://advances.sciencemag.org/content/4/1/eaao5580.

Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway Feedback Loops in Predictive Policing. *arXiv:1706.09847 [cs, stat]*, June 2017. URL http://arxiv.org/abs/1706.09847. arXiv: 1706.09847.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 259–268, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3664-2. doi: 10.1145/2783258.2783311. URL http://doi.acm.org/10.1145/2783258.2783311.

Friedler, Scheidegger, Venkatasubramanian, Choudhary, Hamilton, and Roth. A comparative study of fairness-enhancing interventions in machine learning. In *ACM Conference on Fairness, Accountability and Transparency (FAT*)*. ACM, 2019. URL http://arxiv.org/abs/1802.04422.

Andrew Gelman. What's the most important thing in statistics that's not in the textbooks?, April 2015. URL https://andrewgelman.com/2015/04/28/whats-important-thing-statistics-thats-not-textbooks/.

Abe Gong. Ethics for powerful algorithms (1 of 4), July 2016. URL https://medium.com/@AbeGong/ethics-for-powerful-algorithms-1-of-3-a060054efd84.

Bernard E. Harcourt. Risk as a Proxy for Race. SSRN Scholarly Paper ID 1677654, Social Science Research Network, Rochester, NY, September 2010. URL https://papers.ssrn.com/abstract=1677654.

Moritz Hardt, Eric Price, , and Nati Srebro. Equality of Opportunity in Supervised Learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3315–3323. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf.

Jeff Larson and Julia Angwin. How We Analyzed the COMPAS Recidivism Algorithm, May 2016a. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Jeff Larson and Julia Angwin. Technical Response to Northpointe, July 2016b. URL https://www.propublica.org/article/technical-response-to-northpointe.

Zachary C. Lipton. The Mythos of Model Interpretability. June 2016. URL https://arxiv.org/abs/1606.03490.

Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5):14–19, October 2016. ISSN 17409705. doi: 10.1111/j.1740-9713.2016.00960.x. URL http://doi.wiley.com/10.1111/j.1740-9713.2016.00960.x.

Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*, page 502, San Diego, California, USA, 2011. ACM Press. ISBN 978-1-4503-0813-7. doi: 10.1145/2020408.2020488. URL http://dl.acm.org/citation.cfm?doid=2020408.2020488.

Zehlike Meike, Carlos Castillo, Francesco Bonchi, Mohamed Megahed, Lin Yang, Ricardo Baeza-Yates, and Sara Hajian. Fairness Measures: Datasets and software for detecting algorithmic discrimination, June 2017. URL http://fairness-measures.org/.

Tom Mitchell. *Machine Learning*. McGraw Hill, 1997. ISBN 0-07-042807-7.

Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable.* 2018. URL https://christophm.github.io/interpretable-ml-book/.

Devin G Pope and Justin R Sydnor. Implementing Anti-Discrimination Policies in Statistical Profiling Models. *American Economic Journal: Economic Policy*, 3(3):206–231, August 2011. ISSN 1945-7731, 1945-774X. doi: 10.1257/pol.3.3.206. URL http://pubs.aeaweb.org/doi/10.1257/pol.3.3.206.

Megan Price and Patrick Ball. Big Data, Selection Bias, and the Statistical Patterns of Mortality in Conflict. *SAIS Review of International Affairs*, 34(1):9–20, 2014. ISSN 1945-4724. doi: 10.1353/sais.2014.0010. URL http://muse.jhu.edu/content/crossref/journals/sais_review/v034/34.1.price.html.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016. URL http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf.

Salvatore Ruggieri, Dino Pedreschi, and Franco Turini. DCUBE: Discrimination Discovery in Databases. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 1127–1130, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0032-2. doi: 10.1145/1807167.1807298. URL http://doi.acm.org/10.1145/1807167.1807298.

Jennifer L. Skeem and Christopher Lowenkamp. Risk, Race, & Recidivism: Predictive Bias and Disparate Impact. Technical report, Social Science Research Network, June 2016. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2687339.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. *arXiv:1507.05259 [cs, stat]*, July 2015. URL http://arxiv.org/abs/1507.05259. arXiv: 1507.05259.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. *arXiv:1610.08452 [cs, stat]*, pages 1171–1180, 2017a. doi: 10.1145/3038912.3052660. URL http://arxiv.org/abs/1610.08452. arXiv: 1610.08452.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi, and Adrian Weller. From Parity to Preference-based Notions of Fairness in Classification. *arXiv:1707.00010 [cs, stat]*, June 2017b. URL http://arxiv.org/abs/1707.00010. arXiv: 1707.00010.

Zhe Zhang and Daniel B. Neill. Identifying Significant Predictive Bias in Classifiers. *arXiv:1611.08292 [cs, stat]*, November 2016. URL http://arxiv.org/abs/1611.08292. arXiv: 1611.08292.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. *arXiv:1707.09457 [cs, stat]*, July 2017. URL http://arxiv.org/abs/1707.09457. arXiv: 1707.09457.

Indrė Žliobaitė. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4):1060–1089, July 2017. ISSN 1384-5810, 1573-756X. doi: 10.1007/s10618-017-0506-1. URL http://link.springer.com/10.1007/s10618-017-0506-1.