Fairness: What is the Right Thing to Do? A Comparative Study of Fairness-Preserving Algorithms

Guanqun Yang
Department of Electrical and Computer Engineering
University of California, Los Angeles
Los Angeles, CA 90024
guanqun.yang@engineering.ucla.edu

Abstract

Algorithmic decision making (ADM) system is increasingly more extensively used in people's life, fostering quantitative decision making over subjective judgment. However, since the inherent bias against some underrepresented groups that is deeply rooted inside data, such systems, despite their objectiveness, may also give biased prediction. It is therefore the compliance with fairness requirement in such systems that fairness-preserving algorithms come into play. In this paper, we compare three types of fairness-preserving algorithms that intervene three different phases in machine learning workflow under some fairness metrics. Specifically, uniform sampling in postprocessing phase, fairness regularization in inprocessing phase and rejection option classification in postprocessing phase are compared under type-I parity and type-II parity.

Our comparison shows that all three algorithms could guarantee fairness in the sense of type-I parity or type-II parity. However, in real world application, earliest possible interventions are preferable for most utility.

1 Introduction

Algorithmic decision making (ADM) systems, based on machine learning models, feature their flexibility, accuracy and objectiveness over human decision makers and they are widely applicable in numerous applications including credit trustworthiness evaluation, job applicant selection and automatic essay scoring[1]. However, when it comes to data involving sensitive attributes including race, gender, sex, religion and others, the decisions given by such systems might reflect historical or social prejudices against certain groups and are therefore questionable [2]. FICO score is used to quantify people's credit trustworthiness but it is shown that it is biased against some ethnic groups and therefore it is not considered fair [3]. Amazon reportedly used a résumé selection system that is preferential for male applicants, which exacerbates the low representativeness of female in technology. What is more, automatic essay score used by ETS and many other test services may give will negatively evaluate the essay quality when certain words appear regardless of contents and arguments of essays themselves and therefore bias against the frequent users of that vocabulary.

Facing these issues, many attempts have been made to formalize and foster the fairness in those applications. Dwork et.al. first formalized the fairness in classification and this work serves as the bedrock of many subsequent work[4]. Salerio et.al. provided a comprehensive overview of different notions of fairness and the content each is applicable. They also released an open-source fairness auditing tool that helped visualize potential bias against certain groups. Following the paradigm of three phase workflow of machine learning task, Kamiran et.al. provided methods to address

Preprint. Work in progress.

unfairness in preprocessing phase and postprocessing phase [5, 6]. Kamishima et.al, on the other hand, tried to intervene the machine learning algorithm itself so that fairness is guaranteed [7], which is referred to as inprocessing treatment of unfairness.

Throughout the text, we use y for label and $\widehat{y}$ for predicted label, both takes values in label set $C = \{c+, c^-\}$. The examples $(\mathbf{x}_i, y_i)$ in dataset $\mathcal{D}$ are sampled from underlying distribution $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ may contain some sensitive attributes $A \in \{a_1, a_2, \cdots, a_n\}$ and could therefore be divided into privileged sample domain set $\mathcal{X}^p$ and unprivileged sample domain set $\mathcal{X} \backslash \mathcal{X}^p$.

The paper is organized as follows. Section 1 uncovers the unfairness in algorithmic decision making system through examples and provides a brief review of related work. Section 2 introduces different notions of fairness and algorithms that intervene the machine learning workflow in different phases. Section 3 first visualizes bias in two datasets and then attempts to applies algorithms mentioned in section 2 to relieve unfairness. Section 4 concludes the paper by showing the contribution and further work could be done in related field and disparate perspective.

2 Methods

A machine learning algorithm is said to be fair when predicted outcomes operating on data is non-discriminatory for people based on their protected status such as race, sex, etc[8]. With being non-discriminatory still remaining ambiguous, the characterization of fairness and fairness-preserving algorithm depend on specific machine learning task and therefore result in the taxonomy of different definition to enforce fairness and fairness-preserving algorithms.

2.1 Fairness Metrics

The fairness metrics could be categorized based on the following two principle

- "We Are Equal" (WAE): all groups are similar abilities with respect to the task
- "What You See is What You Get" (WYSIWYG): observations reflect ability with respect to the task.

Take college admission as an example, where standardized test scores like SAT and ACT are the major factor in determining the admission. With WAE principle in mind, groups that are in unfavorable socioeconomic status might feel discriminated since they believe people's potential is not defined by test scores and the gap in these scores result from previous mistreatment to get access to educational resources. However, some other groups, with WYSIWYG principle in mind, might believe test scores are appropriate indicator of ability.

A desirable tradeoff behind these two contradicting principles is type-I and type-II parity [9], which are defined as

- Type-I parity: FDR and FPR

  $$FDR = \Pr[y = c^- | \widehat{y} = c^+, A = a_i]$$
  $$FPR = \Pr[\widehat{y} = c^+ | \mathrm{y} = c^-, A = a_i]$$

  This type of parity is associated with punitive outcomes when the predicted positive.

- Type-II parity: FOR and FNR

  $$FNR = \Pr[\widehat{y} = c^- | y = c^+, A = a_i]$$
  $$FOR = \Pr[y = c^+ | \widehat{y} = c^-, A = a_i]$$

  This type of parity is associated with assistive outcomes when the predicted positive.

2.2 Fairness-Preserving Algorithms

2.2.1 Preprocessing - Uniform Sampling

Suppose $C = \{c^+, c^-\}$, $S = \{p, up\}$, uniform sampling procedure [5] is carried out by first dividing all samples in $\mathcal{D}$ into 4 parts, $\mathcal{D}_{c^+,p}, \mathcal{D}_{c^+,up}, \mathcal{D}_{c^-,p}$ and $\mathcal{D}_{c^-,up}$. Then for each privileged and unprivileged group with label c , the tuple $(s, c), s \in S, c \in C$ shows the distribution of label among groups, where s is determined by sensitive attribute $A = \{a_1, a_2, \cdots, a_m\}$, we compute the weight defined as

$$\begin{aligned} W(s,c) &= \frac{|\{x \in \mathcal{X} : s\}||\{x \in \mathcal{X} : y = c\}|}{|\mathcal{D}||\{x \in \mathcal{X} : s, y = c\}|} \\ &= \frac{\Pr[s]\Pr[y = c]}{\Pr[s, y = c]} \end{aligned}$$

Finally associating weights to each sample in $\mathcal{D}$ and then uniformly select samples from each subgroup of $\mathcal{D}$ with number of samples equal to $W(c^+, p)|\mathcal{D}_{c^+,p}|, W(c^+, up)|\mathcal{D}_{c^+,up}|, W(c^-, p)|\mathcal{D}_{c^-,p}|$ and $W(c^-, up)|\mathcal{D}_{c^-,up}|$, the fairness could be guaranteed.

2.2.2 Inprocessing - Fairness through Regularization

For machine learning algorithms whose loss function is well-defined, an additional fairness regularizer $R(\mathcal{D}, \theta)$ could be added to maximum likelihood estimation formulation of parameter θ [7].

$$-\mathcal{L}(\mathcal{D}; \theta) + \eta \mathrm{R}(\mathcal{D}, \theta) + \frac{\lambda}{2}\|\theta\|_2^2$$

where λ is a tunable hyperparameter.

The fairness regularizer is the KL divergence of $\Pr[y|a_i]$ from $\Pr[y]$, which characterize their difference and is preferably close to zero, showing that a_i is not indicative of y.

$$R(\mathcal{D}, \theta) = \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}} \Pr[y|\mathbf{x}_i; \theta] \ln \frac{\widehat{\Pr}[y|a_i]}{\widehat{\Pr}[y]}$$

In practice $\Pr[y|a_i]$ and $\Pr[y]$ are not directly computable and approximation is used

$$\begin{aligned} \Pr[y|a_i] \approx \widehat{\Pr}[y|a_i] &= \frac{\sum_{(\mathbf{x}_i, y_i) \in \mathcal{D} \text{ s.t. } A = a_i} \Pr[y|\mathbf{x}_i; \theta]}{|\{(\mathbf{x}_i, y_i) \in \mathcal{D} \text{ s.t. } s_i = s\}|} \\ \Pr[y] \approx \widehat{\Pr}[y] &= \frac{\sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}} \Pr[y|\mathbf{x}_i; \theta]}{|\mathcal{D}|} \end{aligned}$$

2.2.3 Postprocessing - Rejection Option Classification

Rejection option classification [6]considers the uncertainty of prediction and defines two cases

- **Critical region**: when probability of prediction is close to 0.5, which shows uncertainty.

  $$\{\mathbf{x} \in \mathcal{X} : \max\{\Pr[c^+|\mathbf{x}], 1 - \Pr[c^+|\mathbf{x}]\} < \theta\},\ 0.5 < \theta < 1$$

- **Stand decision region**: when probability of prediction is high and therefore is certain.

  $$\{\mathbf{x} \in \mathcal{X} : \max\{\Pr[c^+|\mathbf{x}], 1 - \Pr[c^+|\mathbf{x}]\} \geq \theta\},\ 0.5 < \theta < 1$$

  In critical region, we deterministically assign favorable label to unprivileged groups and unfavorable label to privileged group, that is

  $$\begin{aligned} \mathbf{x}_i \in \mathcal{X} \backslash \mathcal{X}^p &\Rightarrow \widehat{y} = c^+ \\ \mathbf{x}_i \in \mathcal{X}^p &\Rightarrow \widehat{y} = c^- \end{aligned}$$

  While in standard decision region, ordinary decision rule is used

  $$\widehat{y} = \underset{\{c^+, c^-\}}{\arg\max}\{\Pr[c^+|\mathbf{x}], \Pr[c^-|\mathbf{x}]\}$$

3 Experiments

The general information of the datasets we use for experiment is shown in Table 1. In student performance dataset, the positive label c^+ is related to favorable outcomes, for example, successfully getting enrolled in college and it is always associated with rewards to these individuals, then type-II parity, i.e. FNR and FOR, is more of interest. That is, we would like to keep the predictions fair among different groups for those who deserve such rewards but fail to because of the prediction error. On the other hand, in adult income dataset, even though the positive label c^+ is generally seen favorable, it could also indicate heavier taxation duties, then type-I parity, i.e. FDP and FPR, is more of interest, where we hope number of people who are falsely responsible for taxation duties are almost the same.

Table 1: General information of experiment datasets

Dataset	Protected Attribute	Target
Student Performance Dataset	Sex	Grade $\geq 60\%$?
Adult Income Dataset	Race, Sex	Salary $\geq$ 50K?

The experimental evaluation follows the diagram shown in Figure 1. We will first show the bias present in the prediction when no interventions are available and then different fairness-preserving algorithms aimed to resolve unfairness will be compared based on metrics we choose. At the same time, in order to control experiments for persuasive comparison, the machine learning algorithm we use is Logistic regression, whose choice is consistent through experiments.

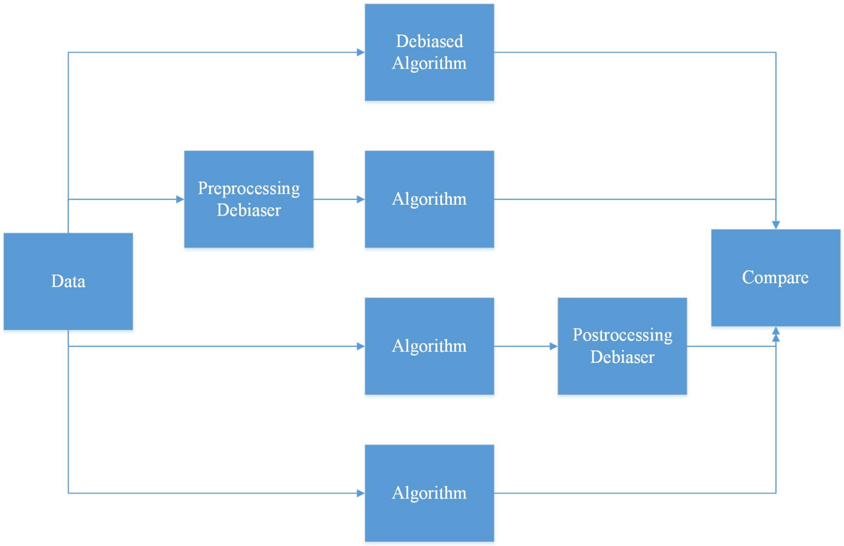

Figure 1: Experimental evaluation procedure

3.1 Student Performance Dataset

3.1.1 Visualizing Bias

As is shown in Figure 2, the female students are biased against when predicting whether a student could pass the exam. Specifically, higher FOR indicates more female students are predicted to fail the exam when they do not and therefore loss the access to rewards.

3.1.2 Removing Bias

As is shown in Figure 3 through Figure 5, the zero-crossing when performing validation shows when specific set of parameter is chosen (in our context, the parameter set is decision threshold of Logistic regression), the FNR and FOR of privileged and unprivileged groups is equal, which indicates fairness.

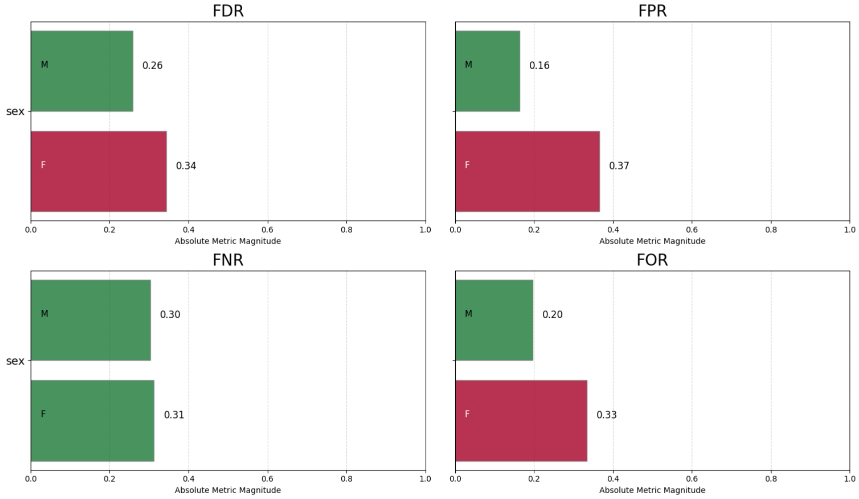

Figure 2: Bias in prediction of student performance dataset

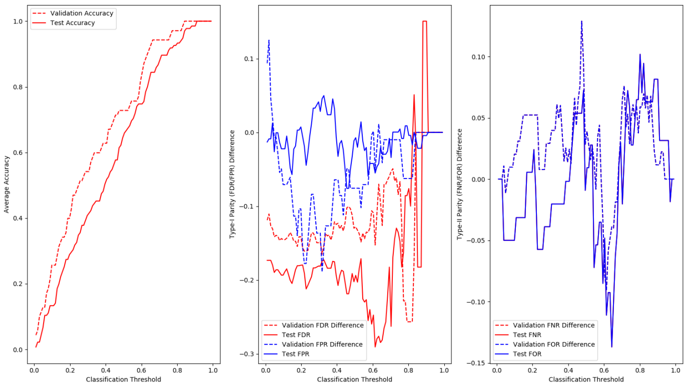

Figure 3: Preprocessing results of student performance dataset

3.2 Adult Income Dataset

3.2.1 Visualizing Bias

As is shown in Figure 6, Black, Asian-Pac-Islander, Amer-India-Eskimo and Female are biased against when predicting he or she could have an annual salary exceeding $50,000. Specifically, higher FDR of these groups will have to be responsible for taxation duties when they should not.

3.2.2 Removing Bias

As is shown in Figure 7 through Figure 9, the zero-crossing when performing validation shows when specific set of parameter is chosen (in our context, the parameter set is decision threshold of Logistic regression), the FDR and FPR of privileged and unprivileged groups is equal, which indicates fairness.

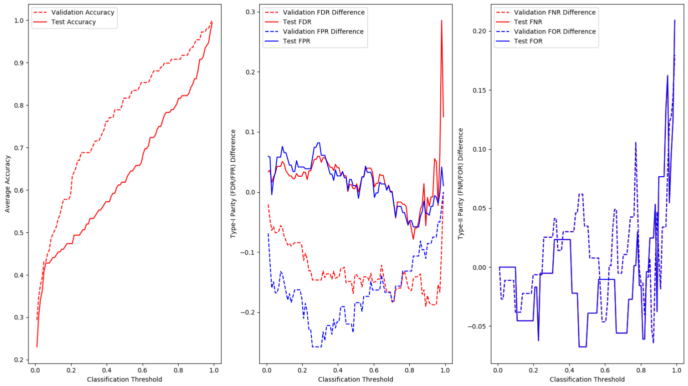

Figure 4: Inprocessing results of student performance dataset

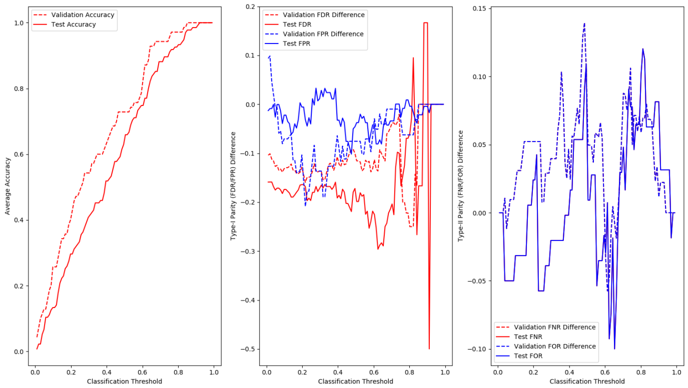

Figure 5: Postprocessing results of student performance dataset

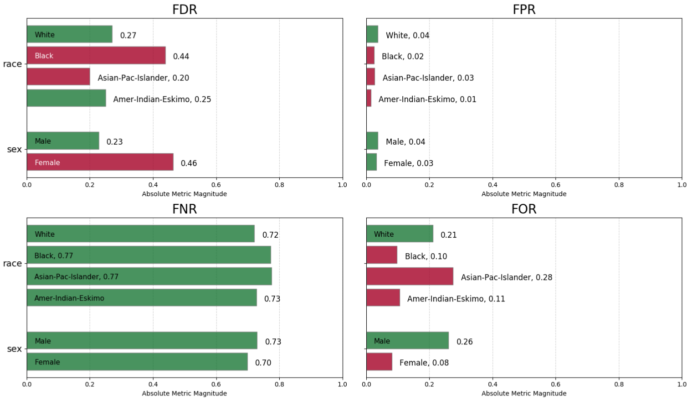

Figure 6: Bias in prediction of adult income dataset

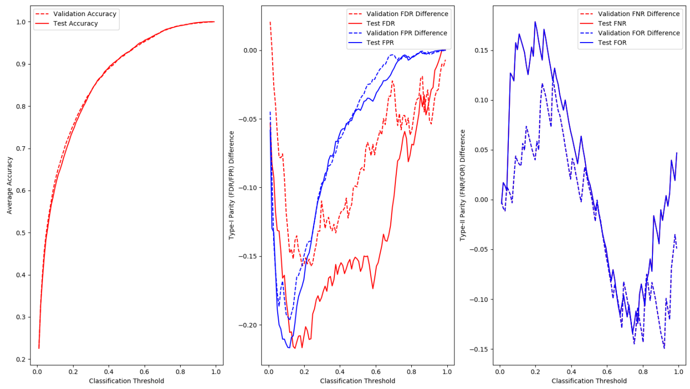

Figure 7: Preprocessing results of adult income dataset

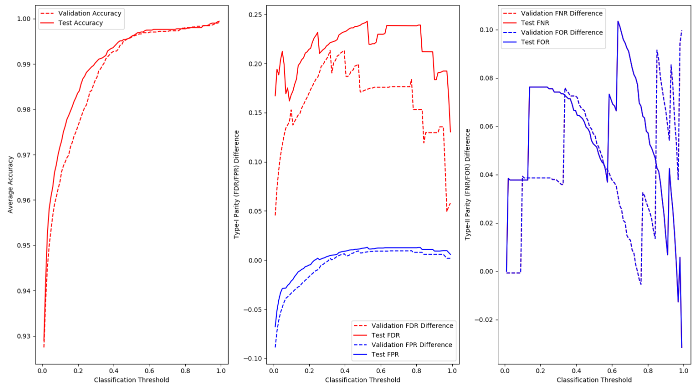

Figure 8: Inprocessing results of adult income dataset

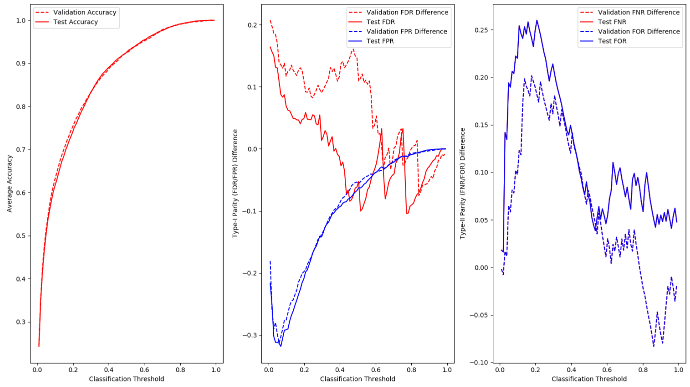

Figure 9: Postprocessing results of adult income dataset

4 Summary and Future Work

The contribution of this paper is two-fold:

- We give an overview of fairness metrics and fairness-preserving algorithms and show difference and applicability among these metrics and algorithms based on scenario the algorithmic decision making system is deployed.
- We compare the performance of three algorithms that intervene different phases of machine learning workflow. Even though fairness is realizable in all three cases, we show that it is preferable to intervene whenever the access to the data and system is possible.

Some future work could be done includes the extension the notion of fairness to deep learning and reinforcement learning, for example, the fairness in face recognition and machine translation, where metrics and algorithms used will be considerably different from our current treatment of classification problem [8]. Additionally, the current statistical methods to enforce fairness is fundamentally flawed in failure to consider the casual structure of sensitive attributes, ordinary attributes and labels. Therefore, the insights from casual reasoning could help, while conceptually not computationally, resolve this limitation [2].

Acknowledgments

We would like to thank Prof. Quanquan Gu for bringing to us this advanced level machine learning course that uncovers the very fundamentals of learning theory, which has long been ignored by us. We would also like to thank Xinzhu Bei, the teaching assistant of this course, for her considerate tutoring, without which no guarantee could be made to survive this course, and she is believed to be the most helpful teaching assistant we have seen. Finally, we would like to acknowledge anyone who is interested in our project and provides their help or shows support for our project.

References

[1] Reuben Binns. Fairness in Machine Learning: Lessons from Political Philosophy. page 11.

[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2018. `http://www.fairmlbook.org`.

[3] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning. *arXiv:1610.02413 [cs]*, October 2016. arXiv: 1610.02413.

[4] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. Fairness Through Awareness. *arXiv:1104.3913 [cs]*, April 2011. arXiv: 1104.3913.

[5] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, October 2012.

[6] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision Theory for Discrimination-Aware Classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929, Brussels, Belgium, December 2012. IEEE.

[7] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-Aware Classifier with Prejudice Remover Regularizer. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Peter A. Flach, Tijl De Bie, and Nello Cristianini, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 7524, pages 35–50. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[8] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. *arXiv:1802.04422 [cs, stat]*, February 2018. arXiv: 1802.04422.

[9] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018.