When data lie: fairness and robustness in contested environments*

Ramya Raghavendra[a], Federico Cerutti[b], and Alun Preece[b]

[a]IBM Thomas J. Watson Research Ctr., USA
[b]Cardiff University, UK

ABSTRACT

Many important decisions historically made by humans are now being made by algorithms—often learnt from data—whose accountability measures and legal standards are far from satisfactory. While model transparency is important, it is neither necessary nor sufficient. Accountability is arguably more important. However, accountability needs to carefully take into consideration the weaknesses of the original data, as well as the weaknesses of the model itself: indeed, robust datasets enable model robustness, and vice versa. In this paper we will focus on unfair datasets, as an example of the weaknesses in datasets. Fairness directly involves privacy problems, since learning without fairness can emphasize certain features or directions that generate private information leakage. For instance, a model may inadvertently reveal a persons age if age is a discriminating feature in a models decision making. Moreover, we will investigate the robustness of model in presence of adversarial activities. Indeed, we should strengthen our models by estimating what an adversary will do based on continuous dynamic learning, mindful of concealment and deception, and with a clear, explainable, insightful summary for the final decision makers. In this paper we will discuss how models based on unfair datasets can hardly be robust; and datasets used by weak models can hardly be fair.

Keywords: machine learning, bias, discrimination, fairness

1. INTRODUCTION

The rise of Machine Learning is every bit as far reaching as the rise of computing itself. A vast new ecosystem of techniques and infrastructures are emerging in the field of machine learning and we are just beginning to learn their full capabilities. But with the exciting things that people can do, there are some really concerning problems arising. Forms of bias, stereotyping and unfair determination are being found in machine vision systems, object recognition models, and in natural language processing and word embeddings.

We are at an inflection point where machine learning is expanding into every area of life from health care, to education to criminal justice. Leading experts in the area envision that the rise of machine learning is every bit as important as the rise of computing itself. Avast new ecosystem of techniques and infrastructures are emerging, and we are just learning their vast capabilities. But amongst the very real excitement of things that ML can do, there is also very real concerns about the problem arising. Forms of bias, discrimination and unfair determination are being found everywhere from machine vision systems, natural language processing and word embedding. Many high profile new stories about bias have made it into main stream media alerting people of the dangers of developing systems without paying attention to systematic biases that they may be learning. Such stories range from women less likely to be shown high paying job advertisements,[1] gender bias in object

*This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

Further author information: (Send correspondence to Ramya Raghavendra)
Ramya Raghavendra: E-mail: ¡rraghav@us.ibm.com¿
Federico Cerutti: E-mail: CeruttiF@cardiff.ac.uk
Alun Preece: E-mail: PreeceAD@cardiff.ac.uk

classification in datasets,[2] racial bias in Google sentiment analysis tool,[3] Amazon same-day delivery system avoiding black neighborhoods,[4] COMPAS scores that show racial bias in criminal justice scores.[5]

The long history of bias lives on in our digital systems, and they become buried into the logic of our machine learning infrastructures. While these are well known stories about bias that made it to the news, these are just the tip of the iceberg while countless backend systems applying off-the-shelf algorithms that can propagate bias in ways that don't have a customer front-end that we not get to see. ML systems are being used by millions of people everyday, so bias matters. This has finally called the attention of industry leaders who are now realizing the real dangers of bias and recognizing that the cost of getting these systems are catastrophic.

Prior work on removing bias from data and models have shown that this is a very hard problem. First of all, the problem is not straightforward since the cause for bias in models comes from bias in training data and we can only gather data about the world we have, which has a long history of discrimination. As such, the default behavior of these systems will be reflect the darkest biases in the society which makes this problem as much technical as a social issue. On the other hand, biased systems that provide incorrect results or keep people unfairly in jail or provide incorrect health care do not serve the machine learning community any better since people would be turned away from using flawed, uninterpretable systems resulting in a decline in interest in the overall field in general.

In the rest of this paper, we will be looking at what is bias and what are its sources followed by which we will be looking at an array of criteria that the community has devised to neutralize the data and scrub models with bias. Next, we look at non-observational methods that actively intervene into the process of data generation and model building. Finally, we summarize the lessons learned and outline the next steps in the direction of addressing the problem of bias.

2. BIAS AND DISCRIMINATION

Bias has a technical meaning in statistical learning, depending on where and how it was introduced. Bias can introduced during data collection including statistical, sampling and reporting bias. An estimator whose expected value differs from the true value of the parameter being estimated is said to be biased. Inductive bias is the set of assumptions that the learner uses to predict outputs given inputs that it has not encountered. While these kind of biases have technical implications as well, it is not this kind of bias that has gained foreground in machine learning discussions lately. The notion of bias that we are concerned with is one that raises ethical issues.

One might be tempted to take a step back and ask the question, *Is the purpose of machine learning not to build classifiers to discriminate?* While it is indeed true that classifiers are built to identify and use discriminating features, the objection is when the basis for that differentiation is unjustified. Further more, such features may be unrelated to the task. For example, a person's race, or gender or sexual orientation would be irrelevant to the task of employment decision. Finally, even if the feature is statistically relevant, it may be morally impermissible to consider certain features. For example, even though hiring a person with disability may be at a higher monetary cost, but we may want to refuse to allow a society in which we confine a person born with the disability to be permanently confined to subordinate positions because they were born with a disability.

It is also important to recognize that discrimination is not a general concept, it is domain specific and is concerned with important opportunities that affect people's life chances. Discrimination is feature-specific and it is concerned with socially salient features that have served as the basis for unjustified and systematically adverse treatment in the past. There are legally regulated domains (incuoding Equal Credit Opportunity Act, Civil Rights Act of 1964; Education Amendments of 1972,Civil Rights Act of 1964,Fair Housing Act,Civil Rights Act of 1964) where private actors are making decisions and bias against legally recognized protected classes (race color, sex, religion, citizenship, age, pregnancy, familial status, disability etc.) is the focus of the work discussed here.

2.1 Discrimination Law

There are two doctrines of discrimination law namely:

- **Disparate treatment:** Disparate treatment is when the bias is intentional. It may be *formal*, where class membership is explicitly considered. Or it may be *intentional*, meaning that discrimination is purposefully attempted without direct reference to class membership but using a clear proxy without considering the feature directly such as the redlining practices followed by the housing industry[6] in the past.

- **Disparate impact:** Disparate impact occurs when a selection process has widely different outcomes for different groups, even as it appears to be neutral. Disparate impact is a more subtle, unintended form of bias in which a decision maker is using features that on the face value seem to have no bearing on the protected attributes, but when used, potentially bears a relationship to a protected class. Ricci v. DeStefano[7] examined the relationship between the notions of disparate treatment and disparate impact, and disparate impact remains a topic of both legal and algorithmic [8] interest.

  In order to prove that a certain decision disparately affects a class, a plaintiff has to first establish perhaps using the four-fifth's rule that a decision procedure has disparate impact. The defendant must provide a justification for making a decision this way, be it necessary for business or job related. Finally, the plaintiff has opportunity to demonstrate an 'alternative practice' wherein the defendant could achieve same goal using a different procedure that would result in a smaller disparity.

Discrimination law aims to achieve procedural fairness and equality of opportunity i.e. all people of equal ability should have equal opportunity. Fairness with disparate impact has a slightly different goal; to find models that minimize the disparate impact that these models may have. Narrow notions of equality of opportunity are concerned with ensuring that decision-making treats similar people similarly on the basis of relevant features, given their current degree of similarity. Broader notions of equality of opportunity are concerned with organizing society in such a way that people of equal talents and ambition can achieve equal outcomes over the course of their lives. Somewhere in between is a notion of equality of opportunity that forces decision-making to treat seemingly dissimilar people similarly, on the belief that their current dissimilarity is the result of past injustice.

2.2 Efforts in Mitigating Bias and Discrimination

Studies have continued to show the incidence and persistence of discrimination through the years. In 2004, there was a study by Bertrand and Mullainathan that showed the callback rate to be 50% higher for applicants with white names than equally qualified applicants with black names.[9] Quillian, Pager et. al. conducted a meta-analysis of the follow-up papers up until 2016 and concluded that there was no meaningful change observed in the degree of discrimination experienced by black job applicants over the past 25 years.[10] As such bias and discrimination remains an incredibly severe and persistent problem.

Interestingly, a lot of the standard approaches to address discrimination and diversity such as bias training and inclusion not only do not mitigate the problem, but empirical studies have shown that it could make the issue worse. What has been known to work is the use of formal procedures in the decision making process that limit opportunities to exercise prejudicial discretion or fall victim to implicit bias. For example, Automated underwriting is shown to increase approval rates for minority and low-income applicants by 30% while improving the overall accuracy of default predictions.[11]

While being a promising path ahead, unfortunately, there is research that suggests that formal procedures can limit opportunities to exercise prejudicial discretion or fall victim to implicit bias. To quote some of the previous work, formal procedures are shown to still leave room for employers to exercise discretion selectively[12] and that bias still affects formal assessments.[13]

2.3 How Machines Learn to Discriminate

In the light of such prior work, a natural question to ask is if Machine Learning (ML) should be viewed as the pinnacle of machine learning. Let's consider a "fair" ML process in which a model would make decisions based on only what the data supports, withhold protected features, and automate decision making, thereby limiting discretion. But there are a lot of problems that make such ideal outcomes really difficult to achieve.[14] We examine a few such issues below.

First, we may have data that comes from a very **skewed sample**. For example, true rate of crime can be very different from reported and observed rate of crimes.[15] Second, we may have **tainted examples**. For example, a model that uses past college admission as a factor in decision making is reflecting on a potentially biased decisions that were made in the past. If one uses annual review scores as a job predictor, then one ends up with a model predicting what the manager *would have* given to a person. Another issue that frequently occurs if that of **limited features**. That is, features may be less informative or less reliably collected for certain parts of the population. A feature set that supports accurate predictions for the majority group may not be accurate for a minority group. Different models with the same reported accuracy can have a very different distribution of error across population.

Even if a feature set is rich and have a perfect representation of the data in the population, by definition, we have fewer samples of a minority group. As a result, we would be able to model the features of the minority group less effectively than a majority group. This is the problem with **sample size disparity**.[16] Another issue is that of the **Proxies** wherein making accurate predictions will mean considering features that are correlated with class membership. With sufficiently rich data, class memberships will be unavoidably encoded across other features.

It important to note that none of the above problems deal with disparate treatment that is formal or intentional, but are accidental, result of inattention or lack of awareness. While machine learning algorithms are not being used to intentionally discriminate. However, disparate impact solutions are not the way forward since the reason discrimination occurs when using machine learning is because "the data says so". The legal process is not well attuned to those issues since the algorithm is reflecting what the data shows. In the end, there are important amount of work to be done so that these problems are acknowledged, characterized and finally addressed.

- We want to be able to discover unobserved differences in performance.
- Even if the data is perfect, we want to be able to cope with observed differences in performance when the model is not going to work as well on some populations as the others.
- We need to push to understand the causes of disparities in predicted outcome.

3. FAIRNESS

3.1 Formal Setup

Consider the running example of a hiring advertisement hiring an Software Engineer (SWE).

A be the features of an individual (browsing history etc.)
X be the sensitive attribute (e.g.: gender) since we are concerned with gender bias in the ad campaign
$C = c(X, A)$ be the predictor (here, whether or not we show an ad)
Y be the target variable. That is, what it is we are trying to predict. (here, SWE)

Formally, these are assumed to be random variables in the same probability space. The conditional probability of an event E, given that we're considering group a is given by:

$$\Pr_a\{E\} = \Pr\{E \mid A = a\} \tag{1}$$

A score function is a random variable $R = r(X, A) \in [0, 1]$. While we have considered a discreet score function, they can very well be continuous, that can easily be turned into a binary predictor by thresholding. In this case, score function is the probability that a user will click on a ad, which is what the predictor is learning. An example of a score function would be a Bayes optimal score which would be optimal for squared loss, given by $r(x, a) = \mathbb{E}[Y \mid X = x, A = a]$.

There has been a surge of interest in defining fairness criteria. Hardt et. al. argue that there are only three fundamental fairness criteria, and all the others are a variant of these. These criteria are:

- **Independence:** C be independent of A.

- **Separation:** C be independent of A conditional on Y.
- **Sufficiency:** Y independent of A conditional on C

In the remainder of this section, we will examine these three fairness criteria and the mathematical framework needed to understand them in order to develop a coherent toolkit to critically examine the many ways that machine learning implicates fairness.

3.2 Independence:

This criteria requires C and A to be independent, denoted by $C \perp A$. That is, for all groups a, b and all values c, $Pr_a\{C = c\} = \Pr_b\{C = c\}$. For a binary predictor, this means that the acceptance rates for the two classes are the same.

3.2.1 Variants of independence:

Sometimes known as the demographic parity, statistical parity, or the 80% rule, it states that when C is binary $Pr_a\{C = 1\} = \Pr_b\{C = 1\}$ for all groups a, b. An approximate constraint can be introduced such that we impose constraints on the ratio of acceptance is one group to acceptance in another group, given by:

$$\frac{Pr_a\{C = 1\}}{\Pr_b\{C = 1\}} \geq 1 - \epsilon \tag{2}$$

The law has a constraint of $epsilon = 20\%$.

3.2.2 Achieving Independence

There has been a number of attempts by the research community on devising ways in which independence can be achieved. Feldman, Friedler et. al. have proposed post-processing methods to transform the input dataset so that predictability of the protected attribute is impossible.[8] We show that this transformation still preserves much of the signal in the unprotected attributes and has nice properties in terms of closeness to the original data distribution.

Calders et. al. in 2009 showed that independence criteria may be achieved by imposing training time constraints including label massaging and reweighing.[17] Another approach that has been explored is pre-processing of the training data i.e. find a representation of your feature space that makes the sensitive attribute independent of that representation.

Pre-processing via representation learning has been proposed by Zemel et. al. [18,19] in which the feature space in which (X, A) are supported. A new representation Z is evolved from this which has some desirable properties (of independence) and we train the classifier on this new representation such that the original feature space is not looked at. The mutual information between $(X; Z)$ is maximized, while the mutual information between $(A; Z)$ are minimized. Lum and Johndrow proposed a method of feature adjustment to remove bias from predictive models by removing all information regarding protected variables from the permitted training data.[20] While this is in no way a comprehensive list of prior work on achieving independence, it provides a good representational set.

3.2.3 Shortcomings of Independence

First, independence as a fairness criteria ignores the possible correlation between target variable Y and sensitive attribute A. In particular, it rules out the perfect predictor $C = Y$ as this would not satisfy the independence criterion in unbalanced datasets. Second, it permits *laziness.* That is, for groups in which there is a good data set, it is possible to create a classifier that works as expected in that it accepts qualified people, but in minority groups without sufficient data it permits randomness. As a result, such a classifier is trading false negatives for false positives by hiding what the errors are, and who is affected by these errors. Finally, it conflates desirable long-term goal with algorithmic constraint. It is not entirely clear that requiring parity in the short term is good for independence in the long run, especially if the short term decisions end up penalizing one demographics in the long term.

3.3 Separation

Separation was introduced as second criterion to overcome some of the shortcomings of Independence, and it attempts to do so by taking the target variable into account. Separation requires the score R is independent of A conditional on target variable Y denoted by $R \perp A \mid Y$. What that means is that for all groups a, b and all values r and y, $\Pr_a\{R = r \mid Y = y\} = \Pr_b\{R = r \mid Y = y\}$. As the reader can observe, it is simply requiring that the conditional independence holds in both the groups. Separation was proposed independently in two separate works by Hardt et. al [21] and by Bilal et. at.[22]

Separation allows for optimality compatibility i.e. $R = Y$ is allowed, which means it allows the the sensitive attribute and target to be correlated as long as it is intrinsic in the target variable. Note that this is not perfect, since we articulated earlier how there could be a lot of bias already accumulated in the target variable. Separation also penalizes laziness, since it requires the same true and false positive rates in all the groups. , Both of these are shortcomings of independence that separation is able to overcome, at least formally.

3.3.1 Achieving Separation

One simple method that has been proposed to achieve separation is post-processing in order to correct the score function.[21] Assuming data about the predictor, target, and membership in the protected group are available, the authors show how to optimally adjust any learned predictor so as to remove discrimination as per their definition. This is done by some thresholding of R (possibly depending on sensitive characteristic A). No retraining or changes to R after training are allowed in this method. For example, the ROC curve of R can be plotted for all possible thresholds and for all groups to obtain the feasible region of tradeoffs, and we can pick an optimal point for a given cost.

Such post-processing would guarantee optimality preservation. That is, if R is close to Bayes optimal, then the output of post-processing is close to optimal among all separated scores. The converse holds, that if we start with a classifier that is not close to optimal, all bets are off. So the alternatives to postprocessing are to collect more data, or if new data is not an option, achieve the constraint at via optimization at training time.[23]

3.4 Sufficiency

The third fairness criterion is sufficiency which states that a random variable R is sufficient for A if $Y \perp A \mid R$. This criterion is called sufficiency since for the purpose of predicting Y (target variable), we do not require to see A (sensitive attribute) when we have R. So, the sensitive attribute, say gender, becomes irrelevant when we have a good classifier for SWE, and we no not need to look at gender as a feature which is very desirable in applications and appealing for legal reasons. Sufficiency is satisfied by Bayes optimal score $r(x, a) = \mathbb{E}[Y \mid X = x, A = a]$.

3.4.1 Achieving Sufficiency

To achieve sufficiency, it is important to realize that sufficiency is implied by *calibration by group* given by $\Pr\{Y = 1 \mid R = r, A = a\} = r$. In each group, the score should gives an actual class probability in each group of a positive outcome, given that score. Calibration by group is a standard technique and can be achieved by various standard calibration methods, for example Platt Scaling.[24,25] Via Platt scaling, given an uncalibrated score R, we would then fit a sigmoid function $S = \frac{1}{1+e^(\alpha R+\beta)}$ of the uncalibrated score against target Y, by minimizing log loss for example.

4. LESSONS LEARNED

Having seen the nature of bias, and some of the attempts in characterizing them and eventually removing them, a natural question to ask next is if these are sufficient in addressing the problem and if not, what remains to be done.

First of all, it is worth noting that satisfying the above three criteria are not only insufficient but actually impossible. It has been shown that any two of the three criteria are mutually exclusive except in degenerate cases and hence tradeoffs are necessary.[26,27] The authors that defined the concept of equal opportunity[21] also created a tool for visualizing trade-offs [†] where it is possible to interactively move the sliders to view why certain

[†] https://research.google.com/bigpicture

situations are not feasible. A significant amount of recent work has followed up these including work on fairness and calibration,[28] optimized pre-processing for discrimination prevention[29] and going beyond parity for fairness objectives for collaborative filtering[30] and preference-based notions of fairness in classification.[31]

Evidently, we are beginning to scratch the surface in understanding how to design fairness criteria and remove discrimination from machine learning algorithms. Note that all the above criteria are *passive* or *observational*, without any what-if scenarios or interventions which leads to inherent limitations.[21] In fact, given that the answer to substantive social questions are not always provided by observational data is in part what motivates the work on causal reasoning and interventions.[32,33] Kusner et.al. have proposed the use of counterfactuals wherein one might questtion: *What would've happened had I been of a different gender when applying to this job?*[33,34]

In conclusion, we make the following observations about fairness and discrimination. First, observational criteria or statistical discrimination criteria can help discover discrimination, but are insufficient on their own. They can point us in the right direction, but they do not provide conclusive proof of (un-)fairness. As a community, instead of establishing the fairness of classifiers it is more productive to engage in what it is that these observational criteria can highlight to us. Next, causal viewpoint can help articulate problems and organize assumptions of the domain and the situation that we are analyzing. Most importantly, social questions starts with measurement. We need to scrutinize features and what they mean more closely. Human scrutiny and expertise irreplacable in understanding what can go wrong. The formalism is a way to guide the human experts, and not to replace them.

Machine learning is domain-specific. We need to understand legal and social context in each domain while designing algorithms. Besides inspecting models, we need to scrutinize data and how it was generated and what assumptions were made. We need to move away from treating ML agorithms as static one-shot problems where we have data and outcome with the goal of minimizing the loss, we need to study the long-term effects of what was intended and how did we achieve that, create feedback loops to understand how the environment is responding, and design interventions to get better data and informative features. We need to establish qualitative understanding of when/why ML is the right tool for the application. For example, do we need to predict crimes, or should be invest resources in preventing crimes through other means.

REFERENCES

[1] Datta, A., Tschantz, M. C., and Datta, A., "Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination," *CoRR* **abs/1408.6491** (2014).

[2] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W., "Men also like shopping: Reducing gender bias amplification using corpus-level constraints," in [*Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*], 2941–2951 (2017).

[3] Lumb, D., "Google's sentiment analysis api is just as biased as humans." https://www.engadget.com/2017/10/25/googles-sentiment-analysis-api-is-just-as-biased-as-humans (2017).

[4] Ingold, D. and Soper, S., "Amazon doesn?t consider the race of its customers. should it?." https://www.bloomberg.com/graphics/2016-amazon-same-day/ (2016).

[5] Angwin, J., Larson, J., Mattu, S., and Kirchner, L., "Machine bias." https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (2016).

[6] "Redlining." https://en.wikipedia.org/wiki/Redlining.

[7] "Ricci v. DeStefano." 557 U.S. 557 (2009).

[8] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S., "Certifying and removing disparate impact," in [*Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*], *KDD '15*, 259–268 (2015).

[9] Bertrand, M. and Mullainathan, S., "Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination," Working Paper 9873, National Bureau of Economic Research (July 2003).

[10] Quillian, L., Pager, D., Hexel, O., and Midtbøen, A. H., "Meta-analysis of field experiments shows no change in racial discrimination in hiring over time," **114**(41), 10870–10875 (2017).

[11] Gates, S. W., Perry, V. G., and Zorn, P. M., "Automated underwriting in mortgage lending: Good news for the underserved?," *Housing Policy Debate* **13**(2), 369–391 (2002).

[12] Wilson, G., Sakura-Lemessy, I., and West, J., "Reaching the top: Racial differences in mobility paths to upper-tier occupations," **26**, 165–186 (1999).

[13] McKay, P. F. and McDaniel, M. A., "A reexamination of black-white mean differences in work performance: More data, more moderators," *Journal of Applied Psychology* **91(3)**, 538–554 (May 2006).

[14] Barocas, S. and Selbst, A. D., "Big data's disparate impact," *104 California Law Review 671* (2016).

[15] Lum, K. and Isaac, W., "To predict and serve?," *Significance* **13**, 14–19 (Obtober 2016).

[16] Hardt, M., "How big data is unfair." https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de (2014).

[17] Calders, T., Kamiran, F., and Pechenizkiy, M., "Building classifiers with independency constraints," in [*2009 IEEE International Conference on Data Mining Workshops*], 13–18 (Dec 2009).

[18] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C., "Learning fair representations," in [*Proceedings of the 30th International Conference on Machine Learning*], **28**, 325–333 (17–19 Jun 2013).

[19] Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R., "The variational fair auto encoder," (11 2015).

[20] Lum, K. and Johndrow, J. E., "A statistical framework for fair predictive algorithms," *CoRR* **abs/1610.08077** (2016).

[21] Hardt, M., Price, E., and Srebro, N., "Equality of opportunity in supervised learning," in [*Proceedings of the 30th International Conference on Neural Information Processing Systems*], 3323–3331 (2016).

[22] Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P., "Fairness beyond disparate treatment &#38; disparate impact: Learning classification without disparate mistreatment," in [*Proceedings of the 26th International Conference on World Wide Web*], *WWW '17*, 1171–1180 (2017).

[23] Woodworth, B. E., Gunasekar, S., Ohannessian, M. I., and Srebro, N., "Learning non-discriminatory predictors," *CoRR* **abs/1702.06081** (2017).

[24] Platt, J. C., "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in [*ADVANCES IN LARGE MARGIN CLASSIFIERS*], 61–74, MIT Press (1999).

[25] Niculescu-Mizil, A. and Caruana, R., "Predicting good probabilities with supervised learning," in [*Proceedings of the 22Nd International Conference on Machine Learning*], *ICML '05*, 625–632 (2005).

[26] Alexandra, C., "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data* **5**(2), 153–163 (2017).

[27] Kleinberg, J. M., Mullainathan, S., and Raghavan, M., "Inherent trade-offs in the fair determination of risk scores," *CoRR* **abs/1609.05807** (2016).

[28] Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. M., and Weinberger, K. Q., "On fairness and calibration," *CoRR* **abs/1709.02012** (2017).

[29] du Pin Calmon, F., Wei, D., Ramamurthy, K. N., and Varshney, K. R., "Optimized data pre-processing for discrimination prevention," *CoRR* **abs/1704.03354** (2017).

[30] Yao, S. and Huang, B., "Beyond parity: Fairness objectives for collaborative filtering," in [*Advances in Neural Information Processing Systems 30*], Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., eds., 2921–2930 (2017).

[31] Zafar, M. B., Valera, I., Gomez-Rodriguez, M., Gummadi, K. P., and Weller, A., "From parity to preference-based notions of fairness in classification," in [*NIPS*], (2017).

[32] Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B., "Avoiding discrimination through causal reasoning," in [*NIPS*], (2017).

[33] Nabi, R. and Shpitser, I., "Fair inference on outcomes," (05 2017).

[34] Russell, C., Kusner, M. J., Loftus, J., and Silva, R., "When worlds collide: Integrating different counterfactual assumptions in fairness," in [*Advances in Neural Information Processing Systems 30*], Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., eds., 6414–6423 (2017).