Algorithmic Fairness: A literature Review

Jonatan M. Contreras

Abstract—The prevalent use of machine learning algorithms in socially-sensitive areas of decision making renders the necessity for fairness definitions, algorithms, and techniques that aid in the intervention and mitigation of bias propagation and unfair classifications of individuals. This emergent field of computer research has been previously referred to as *algorithmic fairness*. Algorithmic fairness has two main components: fairness metrics and bias mitigation algorithms. A fairness metric is a quantification of unwanted bias in training data or models. There are three sets of bias mitigation algorithms: pre-processing, in-processing, and post-processing. Pre-processing algorithms focus on transforming the data into a fairer representation of the data; in-processing algorithms focus on constraining a machine learning algorithm such that it will make fairer classification decisions; and post-processing algorithms focus on post-hoc decision making such that the decisions made become fairer. All three types of algorithms focus on classification tasks.

Index Terms—Algorithmic Fairness, Bias Mitigation Algorithms

INTRODUCTION

Machine learning is ubiquitous. From loan approval decisions [10], to hiring decisions [9], to admissions decisions [25], and even sentencing decisions in prisons [13]; machine learning is found making, or helping humans make, critical and socially-sensitive decisions that affect other people. However, it has been demonstrated that a lot of machine learning algorithms used in these socially-sensitive areas can introduce or propagate biases against groups of people. For example, UT stopped using their GRADE algorithm for admissions because of concerns for equity [4]. ProPublica demonstrated that a risk-assessment tool used to assess the recidivism of criminal defendants (which will in term affect the defendants with parole decisions), COMPAS, would consistently predict Black defendants to be at a higher risk of recidivism than they actually were [21]. A report was published by the Student Borrower Protection Center demonstrating how borrowers can be discriminated against by the machine learning algorithms used in lending [24]. And Amazon's hiring tool demonstrated bias against women [9].

This machine learning dilemma has been identified by researchers since the late 2000's and research attempting to limit this bias mitigation began. Initially, a data mining approach was taken [23], but machine learning researchers began contributing to this field of research as well and established three sets of algorithmic fairness algorithms, each with the goal of increasing the fairness of machine learning decisions. These sets of algorithms are: pre-processing algorithms, in-processing algorithms, and post-processing algorithms [8]. Pre-processing algorithms transform data sets such that they are fairer representations of the individuals in the data set [8]; in-processing algorithms modify or constrain existing machine learning algorithms such that their decision making process during training is fairer [8]; and post-processing algorithms use a holdout set of the data to ensure that the decisions made by the machine learning algorithms are fair, and if not, change some of the decisions made by the machine learning algorithm so the overall set of decisions are fairer [8].

On fairness Metrics

"A fairness metric is a quantification of unwanted bias in training data or models." [3]. Bias mitigation in machine learning is currently modeled as an optimization problem. If we can quantify (un)fairness accurately, then the objective is to optimize for this metric. Thus, the field has developed many fairness metrics. The authors from [22] discuss 10 different fairness metrics. [3] claim they have implemented over 70 fairness metrics in their bias mitigation tool *AI Fairness 360*. The topic of fairness metrics is a vast one and merits its own literature review. However, a few important fairness metrics and concepts are reviewed.

Broadly speaking, fairness metrics in the literature compare the ratio of a given measure between two groups to see if decision processes are fair, the two groups being the majority or privileged group and the minority or under-privileged group. A measure of 0 tends to represent a perfect balance between the groups and the achievement of fairness. To better exemplify this, let's assume a bank wants to use machine learning to make loan approval decisions. The bank will classify individuals into "not approved", or class 0, and "approved",

or class 1. We make no assumptions regarding the data set or machine learning algorithms used as many of these algorithms and fairness metrics are designed to be used broadly across data sets and machine learning algorithms.

To help with clarity, the following notation is adopted from [16]:

Let A be a set of attributes $A = \{a_1, a_2, ..., a_m\}$
Let C be a set of class labels $C = \{c_1, c_2\}$, binary classification is assumed.
A labeled dataset D over A with labels from C is a finite set of tuples $(x_1, ..., x_n, c), c \in C$.
$x.SA = s$ refers to whether the datapoint x has the sensitive attribute of belonging to the under-privileged group
$x.SA \neq s$ refers to belonging to the privileged group.
x^+ refers to a datapoint x being assigned to the positive class in question (e.g., being approved for the loan).
x^- refers to a datapoint x being assigned to the negative class in question (e.g., being denied for the loan).
$D = \{(x^1, c^1), ..., (x^n, c^n)\}$

[16] defines *discrimination* as the difference between the percentage of positive classifications given a data point x belongs to the privileged group and the percentage of positive classifications given a data point x belongs to the under-privileged group. This definition is modeled in equation 1. A positive classification refers to being assigned to the class that has a positive outcome, such as being approved for a loan. In our bank example, a discrimination measure of 0 would mean that the percentage of individuals who belong to a previously defined under-privileged group (in this example, age would be fitting) and are assigned to class 1 is equal to the percentage of individuals who do not belong to a previously defined under-privileged group.

Discrimination:

$$\frac{\{x : \ x \in x^+ \ \wedge \ x.SA \neq s\}}{\{x : x.SA \neq s\}} - \frac{\{x : \ x \in x^+ \ \wedge \ x.SA = s\}}{\{x : \ x.SA = s\}} \tag{1}$$

This fairness metric is widely used in the rule-based literature discussed below [16], [5], [18], [17], [6], as well as the in-processing algorithm described in [19].

Another important fairness metric introduced by [14] is *equality opportunity*. Equality of opportunity states that a machine learning algorithm satisfies equal opportunity if the probability that a given x ∈ D such that x.SA = s is labeled in the positive class 1 is equal to the probability that a given x ∈ D such that x.SA $\not=$ s is labeled in the positive class 1. This is modeled on equation 2:

Equal Opportunity:

$$Pr\{x.c = 1 | x \in D, x.SA = s\} = Pr\{x.c = 1 | x \in D, x.SA \neq s\} \tag{2}$$

This metric dictates that fairness occurs when a machine learning algorithm's predictions are not correlated with the attribute that defines the privileged/unprivileged status of a data point.

[26] borrowed disparate impact from legal literature. Legally, a selection process has disparate impact if that selection process has different outcomes for different groups of people. Informally, the 80% rule is adopted where the ratio between the probability that an under-privileged group of people will be positively classified (e.g., loan approved) and the probability that a privileged group of people will be positively classified is no more than 0.8. [26] models this definition as follows:

Disparate Impact:

$$\frac{Pr\{x.c = 1 | x \in D, x.SA \neq s\}}{Pr\{x.c = 1 | x \in D, x.SA = s\}} \leq 0.8 \tag{3}$$

If this ratio is greater than 0.8, than disparate impact is present in the data set. This threshold is arbitrarily chosen and allows for flexibility depending on the classification task. This is also done to reflect the Supreme Court's definition of disparate impact which resists a "rigid mathematical formula" [2]

There are many more fairness metrics in the literature, but many are defined using this ratio concept. Testing is needed to understand which metric is best for a given situation. The following algorithms will take a fairness metric and optimize it.

Pre-Processing Algorithms

Historical biases can be embedded in data, and thus, training on this data can propagate this bias. Pre-processing algorithms operate on the intuition that classification will be fairer if we simply make the data set fairer. Two general sub-branches of research can be observed in the pre-processing algorithms: transforming the data set given some algorithm (rule-based); and modeling data set transformation as an optimization problem, allowing for the *learning* of a new data set representation that is ultimately fairer than the original data set (learning-based).

Rule-Based Literature

[23] is one of the earliest publications that establishes this precedence by modeling direct and indirect discrimination classification rules and itemsets and using these definitions to then identify

direct/indirect discriminatory rules in the data set. After [23], a lot of seminal work was done by Kamiran and Alders in the rule-based literature between 2009 and 2012 [16], [5], [18] (in-processing technique to be discussed later), [6] (post-processing techinque to be discussed later), [17] (article that compiles and synthesizes work done between 2009-2012).

[16] proposed the classification with no discrimination (CND) algorithm, which transforms the data set. In order to transform the data set, a discrimination measure must be defined. They define discrimination as the difference between the true positive rates between protected and unprotected groups. The closer to 0 this measure is, the more equal these positive rates are, which means that both protected and unprotected groups have an equal opportunity of being classified to the desirable label. This is later formally defined in the literature as *Equal Opportunity* by [14]. The data set is transformed as follows: Two groups are first defined; Candidates for Promotion (CP) and Candidates for Demotion (CD). Members in CP are data points that belong to the sensitive attribute group (e.g. gender, ethnicity, age, etc) whose class is not the desirable one (e.g., "deny loan"), and members in CD are data points that don't belong to the sensitive attribute group whose class is the desirable one. Using Naive Bayes classification, the probability of each data point in the respective groups for belonging to the positive class is computed, based on their frequencies within the data set. Then the members of CP are sorted in descending order in respect to their probabilities and the members of CD are sorted in ascending order in respect to their probabilities. From here, M class labels must be changed in each respective group, where M is the number of class labels needed to be changed in order to bring the measure of discrimination to 0. This assumes binary classification so classes will be changed from 0 to 1, or 1 to 0.

[5] extend [16] in two ways. First [5] is concerned with reducing the dependency between data point attributes and a given class label. Dependency is defined similarly to discrimination in [5]; it is the difference between the true positive rates of data points with a given attribute and the true positive rates of data points without that given attribute. [5] conjectures that simply by reducing this dependency across all attributes will then reduce discrimination. Second, [5] introduces the reweighing method for transforming the data set: instead of ranking the items in each class and changing labels, a weight is computed for data points that have the protected attribute and have a lower true positive rate. The transformed data set is then created by sampling the original data set with replacement according to these new weights. These new weights, then, give the data points that have a dependency between attribute and class a higher probability of being picked than those that don't. This will reduce dependency, eventually to 0, because the true positive rates of the data points with this attribute will increase.

Results were promising for both [16] and [5]. Both methods were successful in reducing discrimination and dependency in the data sets, while maintaining accuracy competitive with classifiers that were trained on non-transformed data sets. However, a notable omission by [5] is that they did not test whether reducing dependency in fact resulted in less discriminating classification results. An observation made in both [16] and [5] is that, whenever a data set is transformed, accuracy will suffer to some degree. This inherent trade-off will be present in all pre-processing algorithms.

Learning-Based Literature

[26] extends the framework by [11] and applies it for pre-processing data. [26] developed a framework that defines a probabilistic mapping from individuals to an intermediate representation such that the mapping achieves fairness by losing information that identifies whether the person belongs to the protected subgroup, while retaining as much of the other information as possible. [11] first posed the data transformation concept as an optimization problem and [26] implemented it by designing an objective function aimed to maintain the new mapping of the data as similar as possible to the original data, while obfuscating protected information, and making the prediction as accurate as possible. There are three terms in the objective function per item mentioned. [26] then minimizes this objective function using L-BFGS.

Their results demonstrated that this method's level of accuracy can compete with Logistic Regression, Naïve Bayes, and Regularized Logistic Regression while demonstrating the least amount of discrimination while doing so. They used the definition of discrimination originally proposed by [16], or equal opportunity. [26] also tested their method against [16]'s method and had better results than [16] on [16]'s metric of discrimination, showing that modeling data transformation as an optimization, or learning problem, as opposed to using rule-based methods, renders better results.

[12] introduced a new fairness metric, disparate impact, to the algorithmic fairness literature as well as a data pre-processing technique that identifies and removes disparate impact. Disparate Impact is a legal term [1] that refers to the occurrence of a selection process having widely different outcomes for different groups, even as it appears to be neutral. Disparate Impact is the predominant legal theory used to determine unintended discrimina-

tion in the US, as opposed to intended or direct discrimination. A data set has disparate impact if the ratio between the likelihood that a data point with a protected attribute is undesirably classified and the likelihood that a data point with a protected attribute is desirably classified is less than or equal to 0.8. In other words, the larger the disparity between these two likelihoods the more disparate impact exists. [12] define disparate impact computationally as the ratio between sensitivity and 1 - specificity; disparate impact exists if this ratio is greater than 1.25. Using this definition, [12] developed two algorithms: one for identifying disparate impact in a dataset, and another for removing it. [12] compared their methods with [16], [20] (an in-processing technique to be discussed later), and [26]. Their discussions of the results is opaque but it's partly due to the combinatorial nature of their data transformation algorithm. Thus, the need for further explorations of these algorithms in a more standardized level becomes more and more apparent as more algorithms are designed and published. However, [12] generally reported competitive results.

[7]'s work is similar to [26] in that they are also presenting an optimization problem with three criteria designated to reduce discrimination. Like [26], [7] aim to 1) reduce discrimination, 2) limit distortion of individual data samples, and 3) maximize accuracy. [7] model their first goal by limiting the dependence of the transformed outcome on any given protected attribute and they use disparate impact to model this dependence; this is very similar to what [12] introduced. The second goal is modeled by applying constraints that reduce or completely avoid large changes from the original data set to the transformed data set. The third goal is modeled by implementing the constraint that the distance between the probability distribution of the original data set and the transformed data set is small. [7] show that, for the most part, this can be considered a convex optimization problem, and thus used a standard convex solver to solve the optimization problem that they modeled. They tested their method against a regularly trained logistic regression and [26]'s method. Although their method was able to reduce the discrimination measure, their accuracy did not outperform [26], demonstrating the limitations of the learning-based approach; once a strong baseline has been established, such as [26], improvements may be substantially difficult to find.

In-Processing Algorithms

The intuition behind in-processing algorithms is that it is possible to constrain machine learning algorithms such that the decisions they make incorporate fairness. One of the first examples in the literature of this is [18]. [18] introduced two techniques for decision tree construction that incorporates fairness in the process; 1) the level of discrimination caused by splitting a tree node is evaluated alongside the accuracy, and 2) during leaf relabeling, the labels of selected leaves are chosen based on its impact on discrimination as well as accuracy. Discrimination is the same measure introduced by [16]. They tested their method against a naive bayes classifier as well as the methods proposed by [16] and [5] and their method outperformed both in terms of accuracy and discrimination.

Around the same time [6] was exploring both in-processing and post-processing techniques using Naive Bayes. The in-processing technique that [6] pioneered for Naive Bayes was to change the observed probabilities in a Naive Bayes model such that its predictions will become discrimination free. This is done by adding probability to the protected attribute given a desirable class. This method is similar to their reweighing technique [5]. In their testing, they demonstrated that their in-processing technique was more successful at removing discrimination and maintaining high accuracy than their post-processing technique.

Although [20] discuss three different sources of discrimination, they develop a regularizer that removes the existence of one of these, which they refer to as *prejudice*. [20] further breaks prejudice down into three sub-definitions, direct, indirect, and latent. In their paper, they focus on minimizing indirect prejudice. Indirect prejudice refers to the statistical dependence between a protected attribute and a class. They quantify indirect prejudice through a measure they refer to as PI, which models the mutual information between the protected attribute and the given class. Even though their regularizer can be implemented into any probabilistic model, they use logistic regression for their implementation. [20] added two regularizers to the logistic regression objective function in order to enforce fairness: the first regularizer was an L2 regularizer to to avoid over fitting, and the second was their fairness regularizer that minimizes their indirect prejudice measure, PI. The authors claim that adding these two regularizers to any probabilistic model's objective function should in turn result in fairer classifications. [20] tested their methods against [6]'s methods and [6]'s method was more successful at reducing discrimination and maintaining a high accuracy. However, [20] demonstrated their methods were still competitive versus regular classifiers without the indirect prejudice regularizer.

More recently, [27] proposed an adversarial learning solution. [27] uses three different fairness metrics in their research to demonstrate that their method is fairness metric-agnostic. That is, given a

fairness metric of choice, their methods will still optimize this fairness metric while maintaining high accuracy. [27] trains a predictor to predict classes given a data set, which contains a protected attribute, and also trains an adversary to learn to predict the protected attributes. At training, the predictor is then trained to maximize its accuracy while minimizing the adversary's ability to predict the protected attributes. The results of their experiments demonstrated that the model minimized whatever fairness metric was being used while maintaining a competitive accuracy.

Post-Processing Algorithms

[19] proposed two methods: Reject Option based Classification (ROC) and Discrimination-Aware ensemble (DAE). [19] state that (ROC) works with any probabilistic classifier while (DAE) works with all classifier ensembles. ROC can be interpreted as a cost-based classification where the cost of misclassifying an instance belonging to a protected attribute is a lot higher than the cost of misclassifying an instance that does not belong to a protected attribute. DAE is an extension of this but for ensembles. The results that [19] report show that these methods outperform existing bias mitigation algorithms of the time, which were mainly the rule-based pre-processing algorithms of Kamiran and Calders [16], [18], [5], [6], [17].

[14] proposed a post-processing technique where they adjust a learned predictor in order to remove discrimination. The two measures they use for discrimination is *equalized odds* and *equal opportunity*. Equalized odds exists of the predicted class and any given attribute are independent. Equal opportunity is equalized odds for the specific case where the attribute is a protected attribute. [14] propose a post-processing technique, where a trained classifier is used as base to derive a predictor with equalized odds or equal opportunity. An optimization algorithm produces this optimal derived predictor. Interestingly, [14] test their method using a case study that predict FICO scores (a bank lending decision) and compute the the profit achieved by each method, as a fraction of the max profit achievable. Equal opportunity is comparable in terms of the profit achieved when compared to a regular classifier, but equalized odds performs quite poorly, suggesting that a bank will not be incentivized to use this type of method. There is no comparisons to any other bias mitigation algorithms.

Discussion and Future Directions

Since the conception of this field, a lot of advancement has been completed. Although we have many different types of fairness metrics, algorithms, and approaches to these algorithms, there are a few common themes that have been discovered. 1) Fairness in machine learning is an optimization problem. Viewing the field from this point of view and cross-referencing it with the vast field of optimization helps us see just how unexplored this field really is. There are many optimization methods and approaches that have been left unused in this field. It would be useful to explore the reasons why, whether it be because all the applicable optimization methods have already been used, or, because many optimization methods are just not applicable. 2) Defining the fairness metric to optimize is just as important as the optimization algorithm. If the fairness metric is not well defined, then our algorithm is moot. This is important because fairness metrics are socially defined and failing to define them properly will, directly or indirectly, help the propagation of bias in machine learning, rather than help amend it. 3) The field of Algorithmic Fairness should be interdisciplinary. Fairness metrics measure something indirectly about individuals, and more often than not, these metrics are more or less arbitrarily defined by computer scientists. It is important that we ask experts in defining social metrics for help in defining these metrics so that we can develop more sophisticated fairness metrics. 4) A lot of work has been completed in this field but how can current practitioners use it in their day-to-day machine learning workflow? Ultimately, it is important that this research becomes useful to machine learning practitioners. However, it is unclear how to extend the research to industry. This is the biggest open question thus far. [3] is a good start for this as they have implemented a fairness tool titled AI Fairness 360. This tool provides libraries for machine learning practitioners to easily implement fairness algorithms into their current machine learning workflow. From a practical point of view, this is a good first step. However, the open research question remain: "What algorithms/fairness metrics should I use, given my situation?" Some work has begun on how to answer this question. [15] for example pioneered a method for developing a baseline across mitigation algorithms, irrespective of their type (i.e., pre-processing, in-processing, post-processing), so that these algorithms can be compared on different data sets.

Finally, more work needs to be done to synthesize and generalize these algorithms. A common theme for example is that, although we have many different fairness metrics, many of them have overlapping definitions. Is there a meta-definition that can encompass most, if not all, of these definitions into one single metric? This would be helpful for practitioners because they would be able to focus on a single metric as opposed to having to worry about multiple. Similarly, a lot of pre-processing

algorithms, for example, transform the data sets from one representation to the next. Is there a meta-algorithm that can encompass all transformations of data sets? These generalizations of our current state of the art will help answer the research question mentioned earlier.

REFERENCES

[1] Griggs v. duke power co., 1971.

[2] Watson v. fort worth bank trust, 1988.

[3] BELLAMY, R. K., DEY, K., HIND, M., HOFFMAN, S. C., HOUDE, S., KANNAN, K., LOHIA, P., MARTINO, J., MEHTA, S., MOJSILOVIC, A., ET AL. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias.

[4] BURKE, L. The death and life of an admissions algorithm. *Inside Highered* (April 2020). Accessed: 2021-11-20.

[5] CALDERS, T., KAMIRAN, F., AND PECHENIZKIY, M. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops* (2009), IEEE, pp. 13–18.

[6] CALDERS, T., AND VERWER, S. Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery 21*, 2 (2010), 277–292.

[7] CALMON, F. P., WEI, D., VINZAMURI, B., RAMAMURTHY, K. N., AND VARSHNEY, K. R. Optimized preprocessing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (2017), pp. 3995–4004.

[8] D'ALESSANDRO, B., O'NEIL, C., AND LAGATTA, T. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data 5*, 2 (2017), 120–134.

[9] DASTIN, J. Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters* (October 2018). Accessed: 2021-11-20.

[10] DIGALAKI, E. The impact of artificial intelligence in the banking sector how ai is being used in 2021. *Business Insider* (January 2021). Accessed: 2021-11-20.

[11] DWORK, C., HARDT, M., PITASSI, T., REINGOLD, O., AND ZEMEL, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (2012), pp. 214–226.

[12] 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* (2015), pp. 259–268.

[13] HAO, K. Ai is sending people to jail—and getting it wrong. *Technology Review* (January 2019). Accessed: 2021-11-20.

[14] HARDT, M., PRICE, E., AND SREBRO, N. Equality of opportunity in supervised learning. *Advances in neural information processing systems 29* (2016), 3315–3323.

[15] HORT, M., ZHANG, J. M., SARRO, F., AND HARMAN, M. Fairea: a model behaviour mutation approach to benchmarking bias mitigation methods. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (2021), pp. 994–1006.

[16] KAMIRAN, F., AND CALDERS, T. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication* (2009), IEEE, pp. 1–6.

[17] KAMIRAN, F., AND CALDERS, T. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems 33*, 1 (2012), 1–33.

[18] KAMIRAN, F., CALDERS, T., AND PECHENIZKIY, M. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining* (2010), IEEE, pp. 869–874.

[19] KAMIRAN, F., KARIM, A., AND ZHANG, X. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining* (2012), IEEE, pp. 924–929.

[20] KAMISHIMA, T., AKAHO, S., ASOH, H., AND SAKUMA, J. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (2012), Springer, pp. 35–50.

[21] LARSON, J., SURYA, M., LAUREN, K., AND JULIAN, A. How we analyzed the compas recidivism algorithm. *ProPublica* (May 2016). Accessed: 2021-11-20.

[22] MEHRABI, N., MORSTATTER, F., SAXENA, N., LERMAN, K., AND GALSTYAN, A. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR) 54*, 6 (2021), 1–35.

[23] PEDRESHI, D., RUGGIERI, S., AND TURINI, F. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (2008), pp. 560–568.

[24] STUDENT BORROWER PROTECTION CENTER. Educational redlining, 2020.

[25] WATERS, A., AND MIIKKULAINEN, R. Grade: Machine learning support for graduate admissions. *Ai Magazine 35*, 1 (2014), 64–64.

[26] ZEMEL, R., WU, Y., SWERSKY, K., PITASSI, T., AND DWORK, C. Learning fair representations. In *International conference on machine learning* (2013), PMLR, pp. 325–333.

[27] ZHANG, B. H., LEMOINE, B., AND MITCHELL, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (2018), pp. 335–340.