FAD: Fairness through Adversarial Discrimination

Alexandra Henzinger[1] and Justin Chen[2]

I. Introduction and Motivation

The performance of machine learning algorithms is being scrutinized more and more using metrics other than just accuracy: as algorithms are applied to numerous and diverse aspects of society, *fairness* in ML has become a central concern in building inclusive and socially responsible technologies. Specifically, machine learning models trained just by maximizing predictive accuracy often reflect and incorporate unwanted biases present in their training data. We consider the case of predicting some label $\hat{y}$ based on features x, where x contains a protected feature z, with respect to which we want our predictions to be unbiased. The applications for this setup are numerous: predicting recidivism scores, assessing criminal risk, making hiring decisions, approving loans, or predicting salary protected from gender, race, or zip code.

In order to augment a machine learning algorithm to be more fair, we must formalize definitions of fairness and develop methods to incorporate fairness into the algorithm, for instance via the loss function or via constraints. The notion of *fairness* itself, i.e. what it means for the output of a machine learning model to be *fair* to different groups or demographics, is not clear-cut. Multiple definitions exist: *demographic parity* requires the prediction $\hat{y}$ and the protected variable z to be independent; *equality of odds* requires the prediction $\hat{y}$ and the protected variable z to be conditionally independent given the true label y (intuitively, all demographics should have equivalent true positive and false positive rates); and *equality of opportunity* requires the prediction $\hat{y}$ and the protected variable z to be conditionally independent given the true label $y = 1$ (intuitively, all demographics should have equivalent true positive rates) [3], [4], [1]. Although formulated differently, all of these conditions of fairness intuitively aim for predictions $\hat{y}$ which are equally *effective* regardless of demographic membership (for various definitions of effectiveness).

We plan to compare the performance of algorithms tailored to these different measures of fairness in high-stakes real-world applications. We use adversarial methods to encode these definitions of fairness into machine learning models: a predictor model optimizes for the target task while an adversarial model jointly optimizes to predict the protected variable from the predictor's output. We combine and compare these adversarial methods with post-processing methods that alter classification thresholds in order to satisfy certain fairness metrics.

We are specifically interested in trade-offs between performance on the target task and the fairness metric. How are changes in methods and models reflected in the results in terms of both accuracy and fairness? Can we retain model performance while jointly pursuing fairness?

II. Related Work

Fairness definitions Lum et al. [5] show that removing the protected variable z from the feature vectors x is not sufficient to debias the given model against z (due to correlations between other features and z). Kleinberg et al. [4] prove that (except in highly-constrained cases) it is not possible to either strictly or approximately satisfy all of calibration within groups (i.e. the model is well-calibrated for each group), balance for the negative class, and balance for the positive class (i.e. equality of odds) at the same time. Similarly, Zhang et al. [7] show that achieving demographic parity and equality of odds are also incongruous goals, i.e. impossible to satisfy simultaneously regardless of the prediction method and the application used.

Adversarial Debiasing Past work has attempted to incorporate demographic parity and equality of odds (i.e. 2 of the 3 definitions of fairness proposed) into machine learning models via adversarial learning techniques. By adding an adversary which penalizes predictions $\hat{y}$ which are biased or skewed based on the protected variable z to the neural network, this setup counteracts unwanted biases present in the training data to make *fair* predictions. This approach is generalizable to any prediction tasks (classification or regression), any predictor and adversary models, any protected variables, and any definition of fairness. Thus, the adversarial model was used by Beutel et al. [1] to achieve equality of opportunity for salary prediction (where the adversary operates on a shared hidden layer), by Zhang et al. [7] to achieve demographic parity and equality of odds for word embeddings and for salary prediction (where the adversary's loss function includes an additional projection term, and where the adversary operates on the output layer of the predictor), and by Wadsworth et al. [6] to achieve demographic parity and equality of odds for recidivism prediction (where the adversary again operates on the output layer of the predictor). Zhang et al. [7] prove the optimality of this adversarial debiasing approach under certain technical assumptions, namely "if the predictor converges, it must converge to a model that satisfies the desired fairness definition".

[1] ahenz@stanford.edu
[2] jyc100@stanford.edu

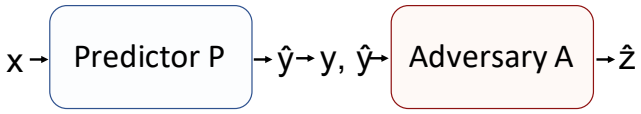

Fig. 1. Adversarial model setup

Thresholds via Post-processing Hardt et al. [3] propose the fairness metric of equality of opportunity (i.e. $\hat{y}$ and z must be conditionally independent given $y = 1$, as presented above), which (similarly to equality of odds) is aligned with the goal of achieving perfect accuracy. Hardt et al. [3] show that "an equalized odds predictor [...] depends on the pointwise minimum ROC curve among different protected groups," and can thus be achieved via a simple post-processing step. As it corresponds to the pointwise minimum, creating such an equalized odds predictor incentivizes models to achieve high accuracy for each demographic (which requires using appropriate features for the prediction task at hand).

III. Methods

A. Adversarial Model

Model structure Let P be our base predictor model and A be our adversary model. The predictor P takes in input features x (including some protected variable z) and outputs a prediction $\hat{y}$. Since we want to satisfy a given fairness definition, we add an adversary to the original predictor which penalizes the original predictor if its prediction $\hat{y}$ is biased against the protected variable z, as shown in figure 1. The inputs for the adversary depend on the fairness metric we are trying to optimize:

- *Demographic Parity Model*: A takes as input the prediction $\hat{y}$ and learns to predict protected variable z. If our prediction is biased against z, we will be able to predict z from $\hat{y}$, and A will have high predictive accuracy.
- *Equality of Odds Model*: A takes as input the prediction $\hat{y}$ and true label y and learns to predict protected variable z. If $\hat{y}$ and z are not conditionally independent given y, the pair $(\hat{y}, y)$ will be indicative of z, and A will have high predictive accuracy.
- *Equality of Opportunity Model*: A takes as input the prediction $\hat{y}$, for all input examples where $y = 1$, and learns to predict protected variable z. If $\hat{y}$ and z are not conditionally independent given $y = 1$, $\hat{y}$ will be predictive of z for examples where $y = 1$, and A will have high predictive accuracy.

Since the model goal is to predict $\hat{y}$ accurately while satisfying some fairness metric, we want predictor P to have a high prediction accuracy and adversary A to have a low prediction accuracy. In this situation, $\hat{y}$ will be a high-accuracy estimate of y, debiased to z, as required.

Model training We use the binary cross-entropy (logistic) loss: let $L_A(y, \hat{y})$ be the logistic loss for adversary A and $L_y(y, \hat{y})$ be the logistic loss for the prediction of $\hat{y}$ in predictor P. We define the loss function for predictor P to be

$$L_P(y, \hat{y}) = L_y(y, \hat{y}) - \alpha L_A(y, \hat{y})$$

in order to pit the predictor and adversary against each other (where α is a hyperparameter that indicates the importance of debiasing $\hat{y}$ according to z). We avoid using the projection term in this loss function as done by Zhang et al. [7] to further simplify the model. After training, the desired condition of fairness should be balanced against predictive accuracy, depending on the the choice of α.

B. Post-training processing

As presented by Hardt et al. [3], we append a simple post-processing step to our adversarial model, which consists of picking the class-specific thresholds t_z for our predictor. Once we select these thresholds (which, without post-processing, we assume to be 0.5), we take as our predicted label $1\{\hat{y}^{(i)} \geq t_{z_i}\}$, for data point i of class z_i. The choice of these thresholds depends on the definition of fairness for which the model is optimizing:

- *Demographic Parity*: demographic parity holds if $\hat{y}$ and z are independent, i.e. $p(\hat{y}|z)$ is equal across all groups z. In post-processing, we pick a fixed value $p(\hat{y})$ (such as to maximize overall accuracy) and then select thresholds t_z such that, for each z, $p(\hat{y}|z) = p(\hat{y})$.
- *Equality of Odds*: equality of odds holds if $\hat{y}$ and z are conditionally independent given y, i.e. the true positive rate (TP) $p(\hat{y} = 1|y = 1, z)$ is equal along all groups z and the false positive rate (FP) $p(\hat{y} = 1|y = 0, z)$ is equal along all groups z. In post-processing, we pick a point along the class-specific ROC curves (which give the TP vs. FP rates for various thresholds) where all class-specific ROC curves intersect (so TP_z is equal for each z and FP_z is equal for each z). Hardt et al. [3] note that it may be necessary to add noise to the predictions made for some classes in order to get the class-specific ROC curves to intersect. This intersection point gives a set of thresholds t_z for each class z.
- *Equality of Opportunity*: equality of opportunity holds if $\hat{y}$ and z are conditionally independent given $y = 1$, i.e. the true-positive rate (TP) $p(\hat{y} = 1|y = 1, z)$ is equal along all groups z. In post-processing, we pick a TP tp^* (i.e. point on the y-axis of the ROC curve) and select the thresholds t_z, for each group z, giving the point where the class-specific ROC curve reaches tp^*.

In each case, we choose the thresholds that maximize predictive accuracy given these constraints. These methods guarantee that the given fairness constraint holds on the dataset used to pick the thresholds (i.e. the validation set).

C. Metrics and Evaluation

For each fairness definition, we use different metrics to evaluate to what extent the fairness condition is achieved:

- *Demographic Parity* : The demographic parity gap for a given class z' is $|p(\hat{y}|z = z') - p(\hat{y}|z \neq z')|$. If demographic parity is satisfied, the demographic parity gap must be 0 for each group z.
- *Equality of Odds*: For a given class z', the true positive gap is $|p(\hat{y} = 1|y = 1, z = z') - p(\hat{y} = 1|y = 1, z \neq z')|$ and the false positive gap is $|p(\hat{y} = 1|y = 0, z = z') - p(\hat{y} =$

$1|y=0, z \neq z')|$. If equality of odds is satisfied, the true positive gap and the false positive gap must be 0 for each group z.

- *Equality of Opportunity*: If equality of opportunity is satisfied, the true positive gap (as defined above) must be 0 for each group z.

IV. Experimental Setup

A. UCI Adult Income Dataset

Salary prediction As this project intends to explore the trade-off between accuracy and fairness, we will experiment with multiple fairness definitions and methods. We chose salary prediction as our application, as it serves as a standard benchmark to compare performance to previous work.

TABLE I
UCI Adult Income Dataset Schema

Feature	Type	Description
age	Cont	Age of the individual
capital gain	Cont	Capital gains via investments
capital loss	Cont	Capital losses via investments
education num	Cont	Max level of education (num)
hours per week	Cont	Hours worked per week
fnlwgt	Cont	Inverse sampling probability
income	Cat	>$50K, ≤$50K
sex	Cat	Female, Male
race	Cat	Racial category
native country	Cat	Country of origin
education	Cat	Max level of education
workclass	Cat	Employer type
occupation	Cat	Occupation type
marital status	Cat	Marital status type
relationship	Cat	Family relationship type

Data and Features We use the UCI adult income data set [2], as done by Zhang et al. and Beutel et al. [7], [1]. Given demographic census data about an individual, the model must predict whether their yearly income is > $50k (classification task). Table I shows the features present for each individual in the UCI data set, their types, and their description.

Pre-processing We pre-process the data as done by Zhang et al. [7]: we discard the feature "fnlwgt" (corresponding to the number of people census takers estimate that a given individual represents), we bucketize age (at the boundaries [18, 25, 30, 35, 40, 45, 50, 55, 60, 65]), and we convert categorical features into one-hot vectors. The dataset comes with separate train and test data. We further split the train data into a training and validation set using an 80%/20% split. The processed data has 117 features with 26,048 examples in the training set, 6,513 examples in the validation set, and 16,281 examples in the test set.

Protected variables The protected variables z used are either sex, race, or age. We experiment with multiple protected variables on the same data set out of interest and as these features have different numbers of buckets: sex can take on 2 values; race can take on 5 values; and age can take on 11 values. We focus our analysis in this paper mainly on sex as a protected variable, since the two-class protected variable yields simpler processing and more easily understandable results and visualization. For the other two protected variables (race and age), we present variance in measures of parity and true and false positive rates, as there are too many gaps to present simply (one gap per value of the protected variable).

Table II shows the distributions of the protected variables in the initial dataset (train, validation, and test). The proportion of high income individuals across groups are not equal for each of these protected variables, thus there is no demographic parity in the underlying dataset. As such, achieving demographic parity is necessarily an incongruent goal from maximizing accuracy on this data set (but this is not the case for equality of odds or opportunity, as perfect accuracy would completely achieve fairness).

TABLE II
Protected Variable Distribution in UCI Income data

Protected Var	Value	Distribution
sex	Male	67%
	Female	33%
race	White	85%
	Black	10%
	Asian-Pac-Islander	3%
	Amer-Indian-Eskimo	1%
	other	1%
age	[0,18)	1%
	[18,25)	16%
	[25,30)	12%
	[30,35)	13%
	[35,40)	13%
	[40,45)	12%
	[45,50)	10%
	[50,55)	8%
	[55,60)	6%
	[60,65)	4%
	65+	4%

B. Experiments

Model Architecture To choose the model architecture for our predictor model, we compare logistic regression and a shallow neural network. The logistic regression model achieves an accuracy of 84.5% on the validation data while the neural network achieves an accuracy of 85.3%. In this dataset and in the context of fairness, joint relationships between demographics seems quite important in making effective and fair predictions of income. The linear model cannot express these joint relationships. Thus, we choose the shallow neural network as our predictor P and our adversary A. After performing a hyperparameter search using

accuracy on the validation set as our target metric, we use a single 10-unit ReLU hidden layer, a learning rate of 10^{-3} for 3,000 iterations, dropout regularization with dropout probability 0.5, and an Adam optimizer for P and A. For binary classification, we use a sigmoid output layer. For multi-class classification (i.e. predicting protected variables race or age in the adversary), we use softmax regression.

Experimental Variables We choose sex, race, and age as our protected variables. For each protected variable, we run one set of experiments in which we debias the predictions against this protected variable, for each adversarial setup (i.e. for each fairness definition setup). Further, for each adversarial setup, we run the model for multiple values of α, where α is the hyperparameter balancing the predictor and the adversary cross-entropy losses. The values of α tested are: 0.1, 1, 10, 100, 300, 500, 700, 900. These values were chosen in order to analyze the spectrum where the adversary seems to play little role in the final model to models where the adversary dominates and predictive accuracy suffers significantly. We present the values of α that achieve reasonable accuracy while producing significant improvements on the fairness metric.

Post-processing We focus on post-processing methods that maximize demographic parity and equality of opportunity. We do not implement methods maximizing equality of odds, as equality of opportunity and odds are very similar metrics and, for this dataset, we care most about the precision of our positive predictions (we care more about misclassifying high-income people as low-income than misclassifying low-income people as high-income).

V. Results

Tables III, V, and IV show the accuracy and fairness performance metrics for experiments with sex, race, and age respectively as protected variables. For each protected variable, we show the performance of our *basic* predictor model (no adversary) and our adversarial model geared towards equality of *opportunity*, equality of *odds*, and demographic *parity* respectively. We present the values achieved setting tradeoff-factor $\alpha = 10$, as this achieved a reasonable balance of accuracy and fairness. As discussed above, a parity gap approching 0 indicates that demographic parity is satisfied; a TP-gap approaching 0 indicates equality of opportunity is satisfied; a FP- and TP-gap approaching 0 indicates equality of odds is satisfied.

Debiasing against sex For sex as a protected variable, we compare our model's performance to that of the shallow neural network used by Beutel et al. [1] and the adversarial logistic regression model used by Zhang et al. [7]. Our basic model outperforms both Beutel and Zhang in terms of accuracy. Further, the opportunity model decreases TP-gap to 0.00; the odds model decreases FP- and TP-gaps to 0.03 each; and the parity model decreases the parity-gap to 0.02. For the opportunity and odds adversarial models, we only suffer accuracy losses of less than 1% with opportunity performing better odds (which makes sense as opportunity is a weaker constraint). This indicates that the tradeoffs between accurate and fair models are not too severe for this dataset and protected variable. The parity model suffers more in accuracy, with a drop of slightly more than 1.5%. In addition, the parity model performs quite poorly in terms of the opportunity and odds metrics. This aligns with our prior knowledge that accuracy and parity are competing metrics for this dataset and that achieving multiple fairness metrics in parallel is not feasible.

We further apply the post-processing step to the basic, parity, and opportunity models, during which we enforce that equality of opportunity or demographic parity must hold on the validation set. For both methods, we see that the basic and special-tailored adversarial models perform almost identically after post-processing. Thus, while we see large gains in adversarial models without post-processing, if we tailor the thresholds class by class, we can get fair and accurate predictions using a basic predictor.

TABLE III

TABLE OF RESULTS - SEX AS PROTECTED VARIABLE

Method	Accuracy	Parity-Gap	FP-Gap	TP-Gap
Basic	**85.66**	0.16	0.06	0.10
Opportunity ($\alpha = 10$)	85.44	0.14	0.47	**0.00**
Odds ($\alpha = 10$)	85.02	0.11	**0.03**	**0.03**
Parity ($\alpha = 10$)	83.92	**0.02**	**0.03**	0.29
Beutel [1]	82.33	0.19	0.15	0.07
Zhang [7]	84.5	–	**0.01**	**0.01**
Basic + *EO*	85.46	0.16	0.06	**0.02**
Opportunity + *EO*	85.45	0.16	0.07	**0.02**
Basic + *DP*	83.63	**0.01**	**0.04**	0.29
Parity + *DP*	83.64	**0.00**	**0.04**	0.31

Debiasing against race and age For race and age as protected variables, there are no results in literature against which we can compare. As each of these protected variables can take on more than 2 values, we chose to report the population variance of $\hat{y}$ to approximate the demographic parity gap, i.e. the variance of $p(\hat{y}|z)$ across all demographics z. Similarly, to approximate the FP- and TP-gaps, we report the population variance of FP and TP. We note smaller changes from model to model (perhaps due to the decreased sensitivity of these metrics), but again see that adding an adversary does not substantially decrease accuracy and decreases var($\hat{y}$) when optimizing of demographic parity, decreases var(TP) when optimizing for equality of opportunity, and decreases var(FP) and var(TP) when optimizing for equality of odds, as compared to the basic model.

VI. Discussion

As the protected variable can only take on 2 values and this case is studied in literature, we limit our discussion to considering sex as a protected variable (we have seen the trends for race and age to be similar).

Equality of opportunity Figure 2 gives the ROC curves for the basic and the opportunity models, presenting the final thresholds chosen with and without the post-processing step to enforce equality of opportunity. Without post-processing, the opportunity model clearly has more similar TP rates

TABLE IV

TABLE OF RESULTS - RACE AS PROTECTED VARIABLE

Method	Accuracy	Var of $\hat{y}$	Var of FP	Var of TP
Basic	**85.57**	0.005	0.000	0.024
Opportunity ($\alpha = 10$)	85.40	0.004	0.000	**0.013**
Odds ($\alpha = 10$)	85.64	0.004	**0.000**	**0.014**
Parity ($\alpha = 10$)	85.59	**0.004**	0.001	0.015

TABLE V

TABLE OF RESULTS - AGE AS PROTECTED VARIABLE

Method	Accuracy	Var of $\hat{y}$	Var of FP	Var of TP
Basic	85.67	0.016	0.002	0.040
Opportunity ($\alpha = 10$)	85.10	0.011	0.001	0.023
Odds ($\alpha = 10$)	84.81	0.008	0.001	**0.009**
Parity ($\alpha = 10$)	82.72	**0.004**	0.001	0.018

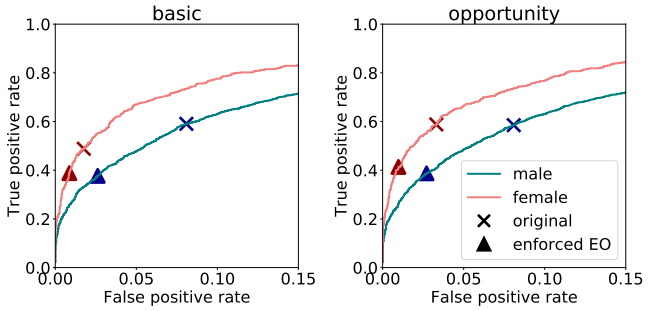

Fig. 2. ROC Curve comparison for basic and opportunity models, with/without post-processing step enforcing equality of opportunity (EO)

between men and women, as desired, since the y-coordinates of the points chosen the ROC curves for female and male are closer. With post-processing, the points along the ROC curves for men and women seem quite similar between the basic and opportunity model. It is interesting to note that the post-processing on both the basic and opportunity model produced a quite different set of predictions (different points on the ROC curve) than the opportunity model by itself even though we see similar results in terms of accuracy and TP-gap.

Demographic parity Figure 3 shows the proportion of positive predictions for men and women as well as the points on the relationship corresponding to the predictions made with and without post-processing for the basic and parity models. We see that the post-processing predictions both fall along the diagonal of the plot, which means that the proportions are equal for men and women, satisfying demographic parity. Those points look to be in the same place for the basic and adversarial models which makes sense given the very similar performance of these two models with post-processing. The original basic predictions are obviously unfair in terms of parity while the original adversarial predictions are very similar to those found with post-processing.

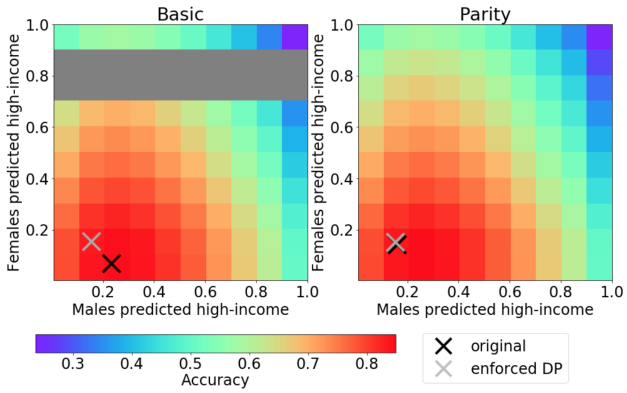

Fig. 3. Accuracy for basic and parity models, with/without post-processing step enforcing demographic parity (DP)

It is interesting to note that while the post-processing methods guarantee satisfying the fairness constraints, the gaps are not quite at zero in the numerical results. This is because the post-processing is done on the validation set and we get the final results from the test set. As this method requires computation of multiple statistics on the validation set, a larger validation set may be required for accurate post-processing as opposed to that needed just for hyperparamter search for the adversarial models. Overall, however, it seems that while adversarial models perform quite well in balancing accuracy and fairness, these results can be matched by these post-processing methods. The adversarial models have the downside of needing extra hyperparameter tuning as well as a more noisy and unpredictable training process while the post-processing methods suffer from a reliance on a large validation set.

VII. CONCLUSION AND FUTURE WORK

To conclude, we explore an combine two ways presented in literature to incorporate fairness metrics into machine learning predictions: we use and adversarial model and a post-processing step analyzing ROC curves to ensure that fairness constraints hold. We perform this task for multiple fairness defintions and multiple protected variables, using the task of predicting an individual's income given census data as the prediction task. We conclude that adversarial methods provide powerful tools to balance between optimization of the target task and preserving fair predictions. Careful post-processing can match these results given a representative validation set even on a basic model that does not factor any notion of fairness into its training.

As future work, it would be interesting to extend this analysis to the multi-class protected variable case and different datasets and protected variables to see if the trends regarding the fairness-accuracy tradeoff, as specified by the ROC curves, still hold. Similarly, testing different predictor and adversary models and architectures (in terms of the structure of the chosen neural network) would again enable us to broaden the analysis conducted here. To further deepen our discussion, we could analyze feature importance in the various models. Finally, incorporating the various fairness definitions discussed into a differentiable loss function would make it possible to take into account fairness constraints in a simple model, without an adversary.

VIII. Contributions and Code

We contributed equally to the project, both designing/writing the code, poster, and report. The code for our project can be found here: https://github.com/jyc8889/fad.

References

[1] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. *CoRR*, abs/1707.00075, 2017.

[2] Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017.

[3] M. Hardt, E. Price, and N. Srebro. Equality of Opportunity in Supervised Learning. *ArXiv e-prints*, October 2016.

[4] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. *ArXiv e-prints*, September 2016.

[5] K. Lum and J. Johndrow. A statistical framework for fair predictive algorithms. *ArXiv e-prints*, October 2016.

[6] Christina Wadsworth, Francesca Vera, and Chris Piech. Achieving fairness through adversarial learning: an application to recidivism prediction. *FAT/ML*, 2018.

[7] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. *CoRR*, abs/1801.07593, 2018.