Towards Explainable and Fair Supervised Learning

Anonymous Authors[1]

Abstract

Algorithms that aid human decision-making may inadvertently discriminate against certain protected groups. We formalize direct discrimination as a direct causal effect of the protected attributes on the decisions, while *induced* indirect discrimination as a change in the influence of non-protected features associated with the protected attributes. The measurements of average treatment effect (ATE) and SHapley Additive exPlanations (SHAP) reveal that state-of-the-art fair learning methods can inadvertently induce indirect discrimination in synthetic and real-world datasets. To inhibit discrimination in algorithmic systems, we propose to nullify the influence of the protected attribute on the output of the system, while preserving the influence of remaining features. To achieve this objective, we introduce a risk minimization method which optimizes for the proposed fairness objective. We show that the method leverages model accuracy and disparity measures.

1. Introduction

Discrimination consists of treating somebody unfavorably because of their membership to a particular group, characterized by a *protected attribute*, such as race or gender. To prevent *disparate treatment*, the law often forbids the use of certain protected attributes, such as race or gender, Z, in decision-making, e.g., in hiring, and dictates that these decisions, Y, shall be based on relevant attributes, $\boldsymbol{X}$, and not depend on the protected attribute, Z. Historically, e.g., in the case of *redlining*, the prohibition of such direct discrimination was sometimes circumvented by the use of attributes correlated with the protected attribute as proxies. This is a particularly acute problem for machine learning data-rich systems, since they often find surprisingly accurate surrogates for protected attributes when a large set of legitimate-looking features is available, resulting in the *inducement* of discrimination via association (Wachter, 2019). To prevent such inducements of discrimination, legal systems establish that the impact of a decision-making process should be the same across groups differing in protected attributes, unless relevant attributes justify it, according to a *business necessity clause* (BNC) (Title VII of the Civil Rights Act, 1964). The main challenge in introducing non-discriminatory learning algorithms lies in preventing the inducement of indirect discrimination, while simultaneously avoiding direct discrimination (Zafar et al., 2015).

Related works. In machine learning, discrimination is typically defined based on statistical independence or causal relations. Well-known fairness objectives, such as parity of impact and equalized odds, correspond to the statistical independence between Z and Y (Hardt et al., 2016; Zafar et al., 2017; Aswani & Olfat, 2019). However, these notions are inconsistent with their legal counterparts (Lipton & Steinhardt, 2019) as legal systems allow for crucial exceptions from this independence through the BNC which permits decisions, Y, to depend on Z through X.

Causal approaches define direct and indirect discrimination as direct and indirect causal influence of Z on Y, respectively (Zhang et al., 2017; Zhang & Bareinboim, 2018; Marx et al., 2019). While this notion of direct discrimination is consistent with the concept of disparate treatment in legal systems, the corresponding indirect discrimination is not consistent with them, since the BNC allows for the use of an attribute that depends on the protected feature (causally or otherwise). This issue is addressed by *path-specific* notions of causal fairness (Nabi et al., 2019; Chiappa, 2019; Wu et al., 2019). These methods allow for *fair causal paths* in which the impact of the protected attribute is permitted, thus, allowing for BNCs. However, if there is no limit on the influence that can pass through such a path, then the path can be used for indirect discrimination, as in the aforementioned case of redlining.

Problem summary. Consider a model supporting human decisions trained on a dataset of n samples $D = \{(\boldsymbol{x}^i, \boldsymbol{z}^i, y^i)\}$, where $\boldsymbol{x}^i \in \mathcal{X}$, $\boldsymbol{z}^i \in \mathcal{Z}$, $y^i \in \mathcal{Y}$, and $i = 1, ..., n$. The goal of a standard supervised learning algorithm is to obtain a function $\hat{y} : \mathcal{X} \rightarrow \mathcal{Y}$ that op-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

timizes a given objective, e.g., $\mathbb{E}[\ell(Y, \hat{y}(\boldsymbol{X}))]$, where the expectation is over the samples in D and ℓ is a loss function. If the dataset is tainted by discrimination, the crucial question is how to drop $\boldsymbol{Z}$ from a model without inducing discrimination, that is without increasing the impact of relevant attributes $\boldsymbol{X}$ correlated with $\boldsymbol{Z}$ in an unjustified and discriminatory way as in redlining.

Contributions. Our work bridges statistical and causal notions of fairness with the literature on explainability, while staying consistent with legal systems. First we define the concepts of direct and *induced indirect discrimination* via measures of causal influence. Second, we construct loss functions, grounded in causality and explainability literature, that measure the change in influence of $\boldsymbol{X}$ while the protected attribute $\boldsymbol{Z}$ is removed. Third, we introduce and evaluate an optimization method that drops the protected $\boldsymbol{Z}$ from a model while minimizing the induction of indirect discrimination through the non-protected features $\boldsymbol{X}$ by minimizing the aforementioned loss functions.

2. Problem formulation

Consider decisions Y that are outcomes of a model that acts on random variables $\boldsymbol{W}$ having support in $\mathcal{W}$. The dimensions of $\boldsymbol{W}$ are indexed, e.g., W_i corresponds to the i'th random variable with support $\mathcal{W}_i$, where $i \in \mathcal{F}$. We distinguish between a set of non-protected attributes $\mathcal{N}$, constituting the $|\mathcal{N}|$-dimensional random variable $\boldsymbol{W}_{\mathcal{N}} = \boldsymbol{X}$, and a set of protected features $\mathcal{P}$, constituting the $|\mathcal{P}|$-dimensional $\boldsymbol{W}_{\mathcal{P}} = \boldsymbol{Z}$. These sets are non-empty, non-overlapping, and the set of all features is $\mathcal{F} = \mathcal{N} \cup \mathcal{P}$.

The model generating decisions Y can suffer the effects of training on discriminatory data. We propose that a non-discriminatory model, $\hat{Y}$, of Y shall remove the influence of the protected features on Y, while preserving the influence of remaining features on Y. In the following subsections, we develop loss functions for supervised learning that aim to achieve this objective.

2.1. Formulation based on causal effect measures

Formal frameworks for causal models include classic potential outcomes (PO) and structural causal models (SCM) (Pearl, 2009) or more recent segregated graphs that include undirected causal relationships (Shpitser, 2015). The methods presented in this work do not rely on the notion of intervention, which tends to have a consistent meaning across causal frameworks.

Note that decisions Y are causal outcomes of the model and the causal parents of these decisions are $\boldsymbol{W}$. This crucial point, emphasized in causal explainability literature (Janzing et al., 2019), allows us to compute influence measures via causal interventions on chosen components of $\boldsymbol{W}$, as if there was no direct causal links between the components of $\boldsymbol{W}$. Following SCM framework, samples of Y are generated by some function, $y = f(\boldsymbol{w}, \epsilon)$, where ϵ is exogenous noise. Since the exogenous noise is unpredictable, here we focus on the de-noised function $y(\boldsymbol{w}) = \mathbb{E}_\epsilon f(\boldsymbol{w}, \epsilon)$. In the notation of SCM and PO, the potential outcome for variable Y after intervention $do(\boldsymbol{Z} = \boldsymbol{z})$ is written as $Y_{\boldsymbol{z}}$. *Average treatment effect* of $\boldsymbol{z}$ on y w.r.t. a reference intervention $\boldsymbol{z}'$ is defined via respective interventions (Pearl et al., 2016),

$$\mathrm{ATE}_Y(\boldsymbol{z}', \boldsymbol{z}) = \mathbb{E}[Y_{\boldsymbol{z}'} - Y_{\boldsymbol{z}}] = \mathbb{E}[Y|\boldsymbol{X}, \boldsymbol{z}'] - \mathbb{E}[Y|\boldsymbol{X}, \boldsymbol{z}],$$

where the last two expectations are over ϵ and a marginal distribution of $P(\boldsymbol{X})$, due to the *causal adjustment* for $\boldsymbol{X}$. A non-causal estimate would use conditional $P(\boldsymbol{X}|\boldsymbol{z})$ instead of $P(\boldsymbol{X})$. The causal *controlled direct effect* of $\boldsymbol{z}$ on y w.r.t. a reference intervention $\boldsymbol{z}'$ and intervention $\boldsymbol{x}$ is

$$\mathrm{CDE}_Y(\boldsymbol{z}', \boldsymbol{z}|\boldsymbol{x}) = \mathbb{E}[Y_{\boldsymbol{x},\boldsymbol{z}'} - Y_{\boldsymbol{x},\boldsymbol{z}}]. \tag{1}$$

Definition 1. *Direct discrimination is the causal influence of a protected attribute $\boldsymbol{Z}$ on the decisions Y in the sense that $\exists_{z,z' \in \mathcal{P}} \exists_{\boldsymbol{x} \in \mathcal{N}} CDE_Y(\boldsymbol{z}, \boldsymbol{z}'|\boldsymbol{x}) \neq 0$.*

To remove this discrimination, one can construct a model $\hat{Y}$ that does not use $\boldsymbol{Z}$. However, this may introduce indirect discrimination into the model via the non-protected attributes X_i associated with the protected attributes $\boldsymbol{Z}$.

Definition 2. *Indirect discrimination induced via X_i is a change in the influence of X_i that depends on $\boldsymbol{Z}$ between the causal process Y and its model $\hat{Y}$, i.e., $\exists_{\boldsymbol{z} \in \mathcal{P}} \exists_{\boldsymbol{x},\boldsymbol{x}' \in \mathcal{N}} CDE_Y(\boldsymbol{x}, \boldsymbol{x}'|\boldsymbol{z}) \neq CDE_{\hat{Y}}(\boldsymbol{x}, \boldsymbol{x}'|\boldsymbol{z})$ such that $P(\boldsymbol{x}|\boldsymbol{z}) \neq P(\boldsymbol{x})$ or $P(\boldsymbol{x}'|\boldsymbol{z}) \neq P(\boldsymbol{x}')$.*

To preserve influence of non-protected attributes we can minimize the following loss

$$L_{\mathrm{ATE}}^{\mathrm{IND}}(\boldsymbol{X}) = \sum_i L_{\mathrm{ATE}}(X_i) =$$

$$\sum_i \mathop{\mathbb{E}}_{X_i'', X_i} \ell(\mathrm{ATE}_Y(X_i, X_i''), \mathrm{ATE}_{\hat{Y}}(X_i, X_i'')).$$

A similar loss could be constructed based on the comparison between $\mathrm{CDE}_Y(\boldsymbol{X}, \boldsymbol{X}''|\boldsymbol{Z})$ and $\mathrm{CDE}_{\hat{Y}}(\boldsymbol{X}, \boldsymbol{X}''|\boldsymbol{Z})$. In this paper we focus on losses that compare ATE and SHAP input influence measures.

2.2. Formulation based on input influence measures

Alternatively, influence can be measured on the grounds of input influence measures introduced to explain black-box AI models.

To measure the influence of a certain variable W_i prior works suggest to make a probabilistic intervention on that variable by replacing it with some W_i' (Datta et al., 2016; Lundberg & Lee, 2017; Janzing et al., 2019). Let the random input variables be $\boldsymbol{W} = \boldsymbol{X}\boldsymbol{Z}$, which are a concatenation of variables $\boldsymbol{X}$ and $\boldsymbol{Z}$. Let primed variables have the same joint distribution as the non-primed variables, $\forall_{\boldsymbol{w}\in\mathcal{W}} P(\boldsymbol{W}' = \boldsymbol{w}) = P(\boldsymbol{W} = \boldsymbol{w})$, while being independent from them, $\boldsymbol{W}' \perp \boldsymbol{W}$. Let double primed variables have the same marginal distributions as the non-primed variables, $\forall_{i\in\mathcal{F}}\forall_{\boldsymbol{w}\in\mathcal{W}_i} P(W_i'' = w) = P(W_i = w)$, and be independent from each other and the non-primed variables, i.e., $\forall_{i\in\mathcal{F}}\forall_{j\neq i} W_i \perp W_j$, $\boldsymbol{W}'' \perp \boldsymbol{W}'$ and $\boldsymbol{W}'' \perp \boldsymbol{W}$. Then, the random variable $\boldsymbol{W}_T\boldsymbol{W}'_{-T} = \boldsymbol{W}_T\boldsymbol{W}'_{\mathcal{F}\setminus T}$ represents a modified random variable $\boldsymbol{W}$ with its components W_i replaced with samples from $P(\boldsymbol{W}')$ for each $i \in \mathcal{F} \setminus T$.

For any subset of features T that does not contain i, we can define a marginal influence (Datta et al., 2016; Janzing et al., 2019)

$$\mathrm{MI}_Y(W_i|\boldsymbol{w}, T) = \mathop{\mathbb{E}}_{\boldsymbol{W}'}\left[Y_{\boldsymbol{w}_{T\cup\{i\}}\boldsymbol{W}'_{-(T\cup\{i\})}} - Y_{\boldsymbol{w}_T\boldsymbol{W}'_{-T}}\right],$$

where $\boldsymbol{W}'$ is a random baseline.

A popular measure of input influence is based on the Shapley value, which averages the marginal influence over all possible subsets T (Datta et al., 2016; Lundberg & Lee, 2017),

$$\mathrm{SHAP}_Y(w_i|\boldsymbol{w}) = \sum_{T\subseteq\mathcal{F}\setminus\{i\}} \frac{\mathrm{MI}_Y(W_i|\boldsymbol{w}, T)}{|\mathcal{F}|\binom{|\mathcal{F}|-1}{|T|}}. \quad (2)$$

To preserve influence of non-protected attributes we can minimize the following loss,

$$L^{\mathrm{IND}}_{\mathrm{SHAP}}(\boldsymbol{X}) = \sum_i L_{\mathrm{SHAP}}(X_i) = \sum_i \mathop{\mathbb{E}}_{\boldsymbol{X}} \ell(\mathop{\mathbb{E}}_{\boldsymbol{Z}''} \mathrm{SHAP}_Y(X_i|\boldsymbol{X}\boldsymbol{Z}''), \mathop{\mathbb{E}}_{\boldsymbol{Z}''} \mathrm{SHAP}_{\hat{Y}}(X_i|\boldsymbol{X}\boldsymbol{Z}'')).$$

3. Minimizing $L_{\mathbf{ATE}}(\boldsymbol{X})$ and $L_{\mathbf{SHAP}}(\boldsymbol{X})$

We seek models $\hat{Y}$ of binary Y that remove the influence of the protected attributes $\boldsymbol{Z}$, while preserving the influence of non-protected attributes $\boldsymbol{X}$ by minimizing $L^{\mathrm{IND}}_{\mathrm{ATE}}(\boldsymbol{X})$ or $L^{\mathrm{IND}}_{\mathrm{SHAP}}(\boldsymbol{X})$ via transfer learning. First, we drop the protected attribute(s) Z from the data. We then obtain "Trad. w/o Z" model by minimizing the cross entropy loss, $H(\hat{y}, y) = -\sum_i y_i \log \hat{y}_i$. Next, we optimize for either $L^{\mathrm{IND}}_{\mathrm{ATE}}(\boldsymbol{X})$ or $L^{\mathrm{IND}}_{\mathrm{SHAP}}(\boldsymbol{X})$. For both objectives we use ℓ_2 loss. We refer to these two-stage optimization-based methods as OPT-ATE and OPT-SHAP, respectively. The training is done using momentum based gradient optimizer ADAM (Kingma & Ba, 2017) via batch gradient descent. We fine-tune two hyper-parameters: learning rate (α) and number of epochs (N). During fine-tuning we pick the values for which we get the best performance on the validation set. In our datasets, α is $1e{-}3$ to $1e{-}2$ and N is from 20 to 100. Our implementations of the methods will be released publicly as a Python library.

4. Experiments

We examine our method's and other supervised learning methods addressing discrimination's performance in binary classification on synthetic and real-world datasets. We measure $\mathbb{E}_{\boldsymbol{X},\boldsymbol{Z}}\,|\mathrm{SHAP}_Y(X_i|\boldsymbol{X},\boldsymbol{Z})|$, following Lundberg & Lee (2017), and $\mathbb{E}_{X_i,X_i'}\,|\mathrm{ATE}_Y(X_i, X_i')|$. To reduce computational costs, we use sub-sampling to compute these. In addition, we measure accuracy, demographic disparity ($|\,P(\hat{y} = 1|z = 0) - P(\hat{y} = 1|z = 1)|$), and equal opportunity difference ($|\,P(\hat{y} = 1|y = 1, z = 0) - P(\hat{y} = 1|y = 1, z = 1)|$). The dataset is partitioned into 20:80 test and train sets. All results are computed on the test set.

4.1. Evaluated learning methods

We evaluate four learning methods addressing discrimination at different stages of a machine learning pipeline (abbreviations in parenthesis). **Pre-processing:** Reweighing approach from Kamiran & Calders (2012). **In-processing:** (1) Reductions model ("Exp Grad") from Agarwal et al. (2018). We evaluate four variations of reductions constraining demographic parity, equalized odds, equal opportunity, and error ratio ("DP", "EO", "TPR", and "ER"). (2) Adversarial debiasing ("Adv Deb") from Zhang et al. (2018). **Post-processing:** Calibrated equalized odds approach ("CalEqOdd') from Pleiss et al. (2017).

We use the implementations of these algorithms as provided in the AI Fairness 360 open-source library (Bellamy et al., 2018). The baseline "traditional" model and underlying classifier for all the evaluated models is logistic regression. We also evaluate a logistic regression model that drops the protected attribute, Z, before training.

4.2. Synthetic results

To generate the synthetic dataset we draw samples from a multivariate normal distribution with standard normal marginals and given correlations. We then convert a column of our matrix into binary values, set that as Z, and set the rest as $\boldsymbol{X}$. The correlations between both (X_1, X_2) and (X_2, Z) are zero. We compare the learning methods while increasing the correlation $r(X_1, Z)$ from 0 to 1. We use a simple model, $Y = \sigma(X_1 + X_2 + Z + 1)$ where σ is the logistic function.

Both OPT approaches preserve X_1's influence with respect to the full model as $r(X_1, Z)$ increases (red and solid blue lines in Figure 1). As expected, the influence of X_1 in-

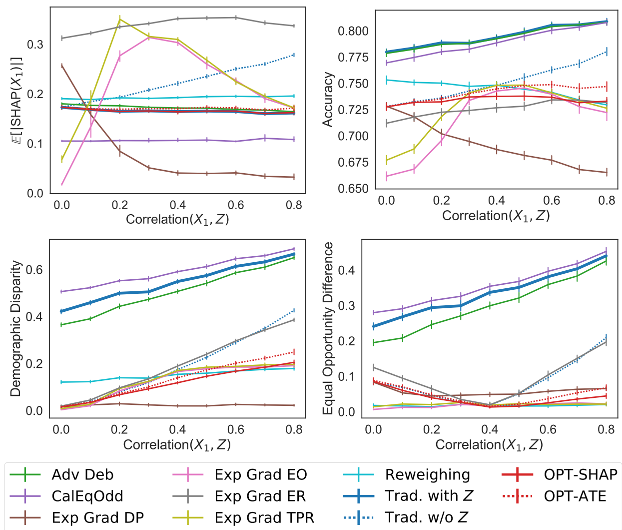

Figure 1. SHAP influence of X_1, model accuracy, and two popular fairness measures as we increase the correlation $r(X_1, Z)$. Error bars show 95% confidence intervals based on 30 samples.

creases with correlation for the traditional method that simply drops Z, i.e., it induces indirect discrimination via X_1 (dotted blue line in Figure 1). Interestingly, even though the OPT does not optimize for either fairness measure, it performs better for all fairness measures than the traditional method dropping Z (in Appendix A we show results for two other fairness measures).

Other methods addressing discrimination either change the influence of X_1 with the growing correlation $r(X_1, Z)$ ("Exp Grad" in Figure 1) or use the protected attribute Z and thus discriminate directly (see "Adv Deb", "CalEqOdd", "Reweighing" in Appendix A). For instance, the method optimizing for parity of impact decreases the impact of $\boldsymbol{X}_1$, because it aims to remove the correlation between $\hat{Y}$ and Z (brown line in Figure 1). Results for ATE are qualitatively the same as for SHAP (Appendix C).

4.3. Real-world results

We train and test the evaluated methods on the COMPAS criminal recidivism dataset (Larson et al., 2016). Here, the model predicts the recidivism of an individual based on their demographics and criminal history with race being the protected attribute. To make the presentation more clear, we exacerbate the racial bias by removing 500 samples of positive outcomes (no recidivism) for African-Americans. Data functions from the AIF360 library are used for this dataset. Results for the unmodified German Credit dataset are qualitatively equivalent (see Appendix B).

In line with the synthetic results, the OPT approaches are not influenced by the protected attribute Z and, with respect to the traditional model, preserve the influence for the two attributes most correlated with Z in this real-world scenario (blue and red in the top row of Figure 2). While most of the evaluated models outperform the OPT models for the fairness measures, they are either influenced by the protected attribute or do not preserve the influence of at least one of the most correlated attributes and have significantly lower accuracy (Appendix A). Therefore, as with the synthetic results, the changes in influence for these attributes indicate that these methods induce indirect discrimination during training, despite having better performance for certain fairness measures.

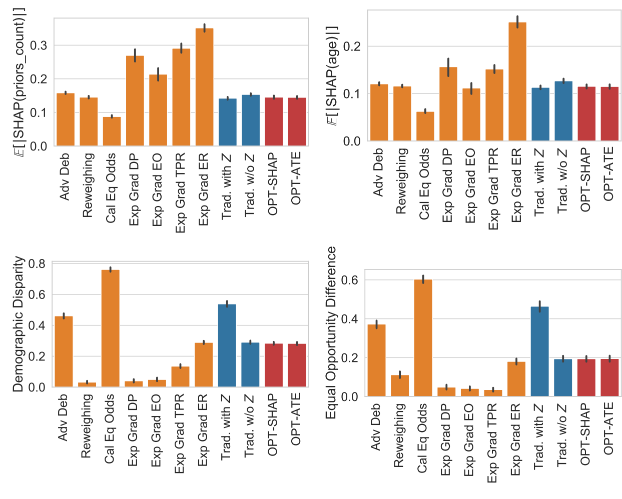

Figure 2. Averaged absolute SHAP for the two features most correlated with the protected attribute and fairness measures for the evaluated models on the COMPAS dataset. Error bars show 95% confidence intervals.

5. Conclusions

The presented results shed a new light on the problem of discrimination prevention in supervised learning. First, we propose a formal definition of induced discrimination, inspired by research in humanist fields (Altman, 2016) and discrimination via association (Wachter, 2019). We measure influence of features to capture induced discrimination. Second, we show that state-of-the-art methods addressing discrimination can return biased models influenced by the protected attribute or attributes associated with it when they are trained on potentially discriminatory datasets. Third, we propose an optimization-based method for discrimination prevention. The method drops the protected attribute and preserves the influence of non-protected attributes to prevent the induction of discrimination via association. These results provide support for the use of the optimization approach in the circumstances where discrimination could have affected the training dataset.

References

Agarwal, A., Beygelzimer, A., Dudfk, M., Langford, J., and Hanna, W. A reductions approach to fair classification. *35th International Conference on Machine Learning, ICML 2018*, 1:102–119, 2018.

Altman, A. Discrimination. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2016 edition, 2016.

Aswani, A. and Olfat, M. Optimization Hierarchy for Fair Statistical Decision Problems. 2019. URL http://arxiv.org/abs/1910.08520.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. October 2018. URL https://arxiv.org/abs/1810.01943.

Chiappa, S. Path-Specific Counterfactual Fairness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:7801–7808, jul 2019. ISSN 2374-3468. doi: 10.1609/aaai.v33i01.33017801. URL https://aaai.org/ojs/index.php/AAAI/article/view/4777.

Datta, A., Sen, S., and Zick, Y. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. *Proceedings - 2016 IEEE Symposium on Security and Privacy, SP 2016*, pp. 598–617, 2016. doi: 10.1109/SP.2016.42.

Hardt, M., Price, E., and Srebro, N. Equality of Opportunity in Supervised Learning. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, pp. 3315–3323. Curran Associates, Inc., oct 2016. ISBN 9781509037612. doi: 10.1109/ICCV.2015.169. URL http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf.

Janzing, D., Minorics, L., and Blöbaum, P. Feature relevance quantification in explainable AI: A causal problem. (2015), oct 2019. URL http://arxiv.org/abs/1910.13413.

Kamiran, F. and Calders, T. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, Oct 2012. ISSN 0219-3116. doi: 10.1007/s10115-011-0463-8. URL https://doi.org/10.1007/s10115-011-0463-8.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2017.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. How We Analyzed the COMPAS Recidivism Algorithm. *Pro Publica*, 2016. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Lipton, Z. C. and Steinhardt, J. Troubling trends in machine-learning scholarship. *Queue*, 17(1):1–15, 2019. ISSN 15427749. doi: 10.1145/3317287.3328534.

Lundberg, S. M. and Lee, S. I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 2017-Decem(Section 2):4766–4775, 2017. ISSN 10495258.

Marx, C. T., Phillips, R. L., Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. Disentangling influence: Using disentangled representations to audit model predictions. *Advances in Neural Information Processing Systems*, 32, 2019. ISSN 10495258.

Nabi, R., Malinsky, D., and Shpitser, I. Learning Optimal Fair Policies. In *Proceedings of the 36th International Conference on Machine Learning*, pp. PMLR 97:4674–4682, sep 2019. URL http://arxiv.org/abs/1809.02244.

Pearl, J. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009. ISBN 0521773628.

Pearl, J., Glymour, M., and Jewell, N. P. *Causal Inference in Statistics: A Primer*. 2016.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On Fairness and Calibration. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5680–5689. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7151-on-fairness-and-calibration.pdf.

Shpitser, I. Segregated graphs and marginals of chain graph models. *Advances in Neural Information Processing Systems*, 2015-Janua:1720–1728, 2015. ISSN 10495258.

Title VII of the Civil Rights Act, 1964. 7, 42 U.S.C., 2000e et seq.

Wachter, S. Affinity Profiling and Discrimination by Association in Online Behavioural Advertising. *SSRN Electronic Journal*, pp. 1–74, 2019. ISSN 1556-5068. doi: 10.2139/ssrn.3388639.

Wu, Y., Zhang, L., Wu, X., and Tong, H. PC-Fairness: A unified framework for measuring causality-based fairness. *Advances in Neural Information Processing Systems*, 32(NeurIPS), 2019. ISSN 10495258.

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. Fairness Constraints: Mechanisms for Fair Classification. *Fairness, Accountability, and Transparency in Machine Learning*, jul 2015. URL http://arxiv.org/abs/1507.05259.

Zafar, M. B., Valera, I., Rodriguez, M. G., Gummadi, K. P., and Weller, A. From Parity to Preference-based Notions of Fairness in Classification. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 229–239. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/6627-from-parity-to-preference-based-notions-of-fairness-in-classification.pdf.

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*, AIES '18, pp. 335–340, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278779. URL https://doi.org/10.1145/3278721.3278779.

Zhang, J. and Bareinboim, E. Fairness in Decision-Making – The Causal Explanation Formula. *AAAI*, pp. 2037–2045, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16949.

Zhang, L., Wu, Y., and Wu, X. A causal framework for discovering and removing direct and indirect discrimination. *IJCAI International Joint Conference on Artificial Intelligence*, 0:3929–3935, 2017. ISSN 10450823. doi: 10.24963/ijcai.2017/549.