Fairness through Equality of Effort

Wen Huang
wenhuang@uark.edu
University of Arkansas

Yongkai Wu
yw009@uark.edu
University of Arkansas

Lu Zhang
lz006@uark.edu
University of Arkansas

Xintao Wu
xintaowu@uark.edu
University of Arkansas

ABSTRACT

Fair machine learning is receiving an increasing attention in machine learning fields. Researchers in fair learning have developed correlation or association-based measures such as demographic disparity, mistreatment disparity, calibration, causal-based measures such as total effect, direct and indirect discrimination, and counterfactual fairness, and fairness notions such as equality of opportunity and equalized odds that consider both decisions in the training data and decisions made by predictive models. In this paper, we develop a new causal-based fairness notation, called equality of effort. Different from existing fairness notions which mainly focus on discovering the disparity of decisions between two groups of individuals, the proposed equality of effort notation helps answer questions like to what extend a legitimate variable should change to make a particular individual achieve a certain outcome level and addresses the concerns whether the efforts made to achieve the same outcome level for individuals from the protected group and that from the unprotected group are different. We develop algorithms for determining whether an individual or a group of individuals is discriminated in terms of equality of effort. We also develop an optimization-based method for removing discriminatory effects from the data if discrimination is detected. We conduct empirical evaluations to compare the equality of effort and existing fairness notion and show the effectiveness of our proposed algorithms.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning*; **Causal reasoning and diagnostics**; • **Applied computing** → **Law, social and behavioral sciences**.

KEYWORDS

Fairness, Equality of Effort, Causality

ACM Reference Format:
Wen Huang, Yongkai Wu, Lu Zhang, and Xintao Wu. 2020. Fairness through Equality of Effort. In *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3366424.3383558

WWW '20 Companion, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7024-0/20/04.
https://doi.org/10.1145/3366424.3383558

1 INTRODUCTION

Fair machine learning is receiving an increasing attention in machine learning fields. Discrimination is unfair treatment towards individuals based on the group to which they are perceived to belong. The first endeavor of the research community to achieve fairness is developing correlation or association-based measures, including demographic disparity (e.g., risk difference), mistreatment disparity, calibration, etc. [3, 4, 14, 19, 36], which mainly focus on discovering the disparity of certain statistical metrics between two groups of individuals. However, as paid increasing attention recently [9, 10, 12, 16, 25, 30, 31, 34, 35], unlawful discrimination is a causal connection between the challenged decision and a protected characteristic, which cannot be captured by simple correlation or association concepts. To address this limitation, causal-based fairness measures have been proposed, including total effect [29], direct and indirect discrimination [29, 34], counterfactual fairness [11, 21, 25], and path-specific counterfactual fairness [2, 26]. Fairness notions have also been extended to considering both decisions in the training data and decisions made by predictive models, such as equality of opportunity and equalized odds [5, 27], and counterfactual direct and indirect error rates [28].

In this paper, we develop a new causal-based fairness notation, called equality of effort. Consider a dataset with N individuals with attributes $(S, T, \mathbf{X}, Y)$ where S denotes a protected attribute such as *gender* with domain values $\{s^+, s^-\}$, Y denotes a decision attribute such as *loan* with domain values $\{y^+, y^-\}$, T denotes a legitimate attribute such as *credit score*, and $\mathbf{X}$ denotes a set of covariates. For a particular applicant i in the dataset with profile $(S_i = s^-, T_i = t, \mathbf{X}_i = \mathbf{x}, Y_i = y^-)$, she may ask the counterfactual question, how much her credit score she should improve such that the probability of her loan application approval is above a threshold γ (e.g., 80%). Informally speaking, our proposed equality of effort notation addresses her concern on whether her future effort (the increase of her credit score) has no difference from male applicants with similar profile $\mathbf{x}$.

Following Rubin's causal modeling notations, we use $Y_i(t)$ to represent the potential outcome for individual i given a new treatment $T = t$, $\mathbb{E}[Y_i(t)]$ to denote the -level expectation of outcome variable. If $\mathbb{E}[Y_i(t)] \geq \gamma$, we say applicant i tends to receive loan approval with at least probability γ. We can then calculate or estimate the minimum value of the treatment variable to achieve γ-level outcome for individual i. If the minimum value of individual i is significantly higher than her counterparts (i.e., males with similar characteristics), discrimination exists in terms of effort discrepancy.

Our fairness notation, equality of effort, is different from existing fairness notions, e.g., statistical disparity, path-specific effects, which mainly focus on the the effect of the sensitive attribute S on the decision attribute Y. Our proposed equality of effort instead focuses on to what extend the treatment variable T should change to make the individual achieve a certain outcome level. This notation addresses the concerns whether the efforts that would need to make to achieve the same outcome level for individuals from the protected group and the efforts from the unprotected group are different. We develop algorithms for determining whether an individual or a group of individuals are discriminated in terms of equality of effort based on three widely used techniques for causal inference, outcome regression, propensity score weighting, and structural causal modeling. We also develop an optimization-based method for removing discriminatory efforts from biased datasets. We conduct empirical evaluations to compare the equality of effort and existing fairness notions and evaluation results show the effectiveness of our proposed algorithms.

2 PRELIMINARIES

2.1 Notations

In this paper, an uppercase denotes a variable, e.g., S; a bold uppercase denotes a set of variables, e.g., $\mathbf{X}$; a lowercase denotes a value or a set of values of the variables, e.g., s and $\mathbf{x}$; and a lowercase with superscript denotes a particular value, e.g., s^+ and x^-.

2.2 Potential Outcomes Framework

The potential outcomes framework, also known as Neyman-Rubin potential outcomes or Rubin causal model, has been widely used in many research areas to perform causal inference. It refers to the outcomes one would see under each treatment option. Let Y be the outcome variable, T be the binary or multiple valued ordinal treatment variable, and $\mathbf{X}$ be the pre-treatment variables (covariates). $Y_i(t)$ represents the potential outcome for individual i given treatment level $T = t$ and $\mathbb{E}[Y_i(t)]$ denotes the individual-level expectation of outcome variable. The "fundamental problem of causal inference" claims that one can never observe all the potential outcomes for any individual [7] and we need to compare potential outcomes and make inference from observed data. We use $\mathbb{E}[Y(t)]$ to denote population-level expectation of outcome variable and $\mathbb{E}[Y_\diamond(t)]$ to denote the conditional expectation of outcome variable within certain sub-population $\diamond$.

Classic causal inference focuses on estimating the potential outcome and treatment effect given the information of treatment variable and pre-treatment variables [1]. For example, the average treatment effect $ATE = \mathbb{E}[Y(t') - Y(t)]$ answers the question of how, on average, the outcome of interest Y would change if everyone in the population of interest had been assigned to a particular treatment t' relative to if they had received another treatment t. The average treatment effect on the treated, $ATT = \mathbb{E}[Y(t') - Y(t)|T = t]$ is about how the average outcome would change if everyone who received one particular treatment t had instead received another treatment t'.

The potential outcome framework relies on three assumptions: (1) Stable Unit Treatment Value Assumption (SUTVA) which basically requires the potential outcome observation on one unit should be unaffected by the particular assignment of treatments to the other units. (2) Consistency assumption which means that the value of potential outcomes would not change no matter how the treatment is observed or assigned through an intervention. (3) Strong ignorability (unconfoundedness) assumption which is equal to the assumption that there are no unobserved confounders. A confounder is a pre-treatment variable that affects both treatment and outcome variables. In this paper, we follow these three assumptions.

2.3 Propensity Score Method

Definition 2.1 (Propensity Score). For a binary treatment variable, propensity score is the conditional probability of receiving treatment T given the pre-treatment variables $\mathbf{X}$,

$$e(\mathbf{x}) = Pr(T = 1|\mathbf{X} = \mathbf{x})$$

The estimation of propensity scores requires the model or functional form of $e(\cdot)$ and the variables to include in $\mathbf{X}$. Let $e(i)$ denote the propensity score for individual i, for binary valued groups, the propensity score is estimated by logistic regression:

$$logit(e(i)) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k,$$

where $x_1, ..., x_k$ are values of the selected covariates and $\beta_1, ..., \beta_k$ are regression coefficients. If correctly estimated, the reciprocal of propensity score can be used as the weight for each individual such that the distribution of the group under treatment 1 and that under treatment 0 becomes identical. [20] showed that conditional on the propensity score, all observed covariates are independent of treatment assignment, and they will not confound estimated treatment effects.

Hence after weighting procedure, a pseudo-balanced population can be built in which the imbalance caused by measured covariates between the treatment groups has been eliminated. The average potential outcome can thus be estimated by some standard estimators. For example, one unbiased estimator of the population-level ATE can be written as: $\frac{1}{N_1}\sum_{i\in N}\mathbb{1}_{T_i=1}\omega_i y_i - \frac{1}{N_2}\sum_{i\in N}\mathbb{1}_{T_i=0}\omega_i y_i$ where $N_1 = \sum_{i\in N}\mathbb{1}_{T_i=1}$ and $N_2 = \sum_{i\in N}\mathbb{1}_{T_i=0}$.

3 FAIRNESS THROUGH EQUAL EFFORT

For the sake of simplicity, we assume there is only one binary protected attribute, one binary decision attribute, and one ordered multi-categorical legitimate attribute. Our formulation and methods are readily to extend to general cases where there are multiple protected/decision/legitimate attributes. In this paper, we simply use the change of T as the effort needed to achieve a certain level of outcome and do not consider the real monetary or resource cost behind that change.

3.1 Equality of Effort at the Individual Level

For an individual i in the dataset with profile $(s_i, t_i, \mathbf{x}_i, y_i)$, we want to figure out what is the minimal change on treatment variable T to achieve a certain outcome level based on observational data. If the minimal change for individual i has no difference from that of counterparts (individuals with similar profiles except the sensitive attribute), we say individual i achieves fairness in terms of equality of effort.

Formally, we use $Y_i(t)$ to represent the potential outcome for individual i given a new or counterfactual treatment $T = t$. We

use $\mathbb{E}[Y_i(t)]$ to denote the individual-level expectation of outcome variable where $\mathbb{E}[\cdot]$ is the expectation operator from probability theory. When $\mathbb{E}[Y_i(t)]$ is larger than a predefined threshold γ, we say individual i would receive a positive decision with probability γ.

Definition 3.1 (γ-Minimum Effort). For individual i with value $(s_i, t_i, \mathbf{x}_i, y_i)$, the minimum value of the treatment variable to achieve γ-level outcome is defined as:

$$\Psi_i(\gamma) = \underset{t \in T}{\operatorname{argmin}} \left\{\mathbb{E}[Y_i(t)] \geq \gamma)\right\}$$

and the minimum effort to achieve γ-level outcome is $\Psi_i(\gamma) - t_i$.

However $Y_i(t)$ cannot be directly observed and we have to derive its estimate from samples with similar characteristics. We design an estimation procedure based on the idea of situation testing [14], which is one normal practice of determining whether an individual is discriminated. How to select variables for finding similar individuals has been studied in situation testing based individual discrimination discovery [32]. The proposed idea there was to first construct a causal graph for all variables and then select variables that are the parents of the decision. Their work is also applicable to our equal effort definition. We first find a subset of users, denoted as I, each of whom has the same (or similar) characteristics ($\mathbf{x}$ and t) as individual i. We denote I^+ (I^-) the subgroup of users in I with the sensitive attribute value s^+ (s^-). Similarly, $\mathbb{E}[Y_{I^+}(t)]$ denotes the expected outcome under treatment t for the subgroup I^+. The minimal effort needed to achieve γ level of outcome variable within the subgroup I^+ is then defined as:

$$\Psi_{I^+}(\gamma) = \underset{t \in T}{\operatorname{argmin}} \{\mathbb{E}[Y_{I^+}(t)] \geq \gamma\}.$$

Definition 3.2 (γ-Equal Effort Fairness at the Individual Level). For a certain outcome level γ, we define equality of effort for individual i if

$$\Psi_{I^+}(\gamma) = \Psi_{I^-}(\gamma).$$

The difference $\delta_i(\gamma) = \Psi_{I^+}(\gamma) - \Psi_{I^-}(\gamma)$ measures the effort discrepancy at the individual level.

3.2 Equality of Effort at the Group or System Level

In addition to the task of checking individual level discrimination, we also want to check whether discrimination exists at the group or system level. System-level discrimination deals with the average discrimination across the whole system, e.g., all applicants to a university, and group-level discrimination deals with discrimination that occurs in one particular subgroup, e.g., the applicants applying for a particular major. Existing works [34, 36] apply demographic disparity metrics (e.g., risk difference) or causal effect (e.g., direct and indirect causal discrimination) on the whole dataset (the subset of data) to determine the system-level (group-level) discrimination. Similarly, we may want to check whether there are effort discrepancies at the group or system level.

We denote D as the whole dataset, and D^+ (D^-) as the subset with the sensitive attribute value s^+ (s^-). We define the minimum value of treatment variable to achieve a certain outcome level γ for D^* as:

$$\Psi_{D^*}(\gamma) = \underset{t \in T}{\operatorname{argmin}} \left\{\mathbb{E}[Y_{D^*}(t)] \geq \gamma\right\}.$$

Definition 3.3 (γ-Equality of Effort at the System Level). For a certain outcome level γ, equality of effort between two sensitive attributes s^+ and s^- is achieved if

$$\Psi_{D^+}(\gamma) = \Psi_{D^-}(\gamma).$$

The difference $\delta_D(\gamma) = \Psi_{D^+}(\gamma) - \Psi_{D^-}(\gamma)$ measures the effort discrepancy at the system level.

Definition 3.3 can be straightforwardly adapted to the group level. Given two compared groups, their distributions in terms of certain attributes (e.g., outstanding debt) could be different. The simple use of our group equal-effort fairness may not be appropriate. In this case, we could apply the path-specific effect/mediator analysis [16, 34] to separate and measure different causal effects e.g., direct discrimination, indirect discrimination, and explainable effects.

3.3 Comparison with Other Fairness Metrics

Many different fairness metrics have been proposed to measure fairness of data and machine learning algorithms. Classic metrics include individual fairness, demographic parity, equality of opportunity, calibration, causal fairness, and counterfactual causal fairness. Refer to a recent survey [24]. We show in Table 1 the formula of previous representative fairness metrics to compare with our equality of effort notion. For example, demographic imparity requires that $P(y^+|s^+) = P(y^+|s^-)$ and similarly conditional demographic imparity requires $P(y^+|s^+, \mathbf{o}) = P(y^+|s^-, \mathbf{o})$ where $\mathbf{o}$ is the values of a specified variable set $\mathbf{O}$. Basically they require that a decision be independent of the protected attribute conditional or unconditional on some other variables. For causal based fairness notions, the total causal discrimination is based on the average causal effect of S on Y and is defined as $\mathbb{E}[Y(s^+)] - \mathbb{E}[Y(s^-)]$, which represents the expected change of outcome Y when S of all individuals changes from s^- to s^+. Different from the total causal discrimination that measures the causal effect transmitted along all the causal paths from S to Y in the causal graph, the path-specific causal discrimination is based on the causal effect that is transmitted along some specific paths π from S to Y, e.g., direct causal discrimination when π is the direct path from S to Y, and indirect causal discrimination when π is all paths from S to Y through redlining attribute T. Counterfactual fairness requires $\mathbb{E}[Y_{\mathbf{o}}(s^+)] - \mathbb{E}[Y_{\mathbf{o}}(s^-)]$, which means that a decision is fair towards an individual if it is the same in the actual world and a counterfactual world where the individual belonged to a different demographic group. Most recently, [26] developed a unified definition, path-specific counterfactual fairness (PC Fairness), that covers previous causality-based fairness notations. Different from demographic parity and causal based fairness notions, our proposed equality of effort considers to what extend the legitimate variable T should change to achieve a certain outcome level and whether the minimum effort made for individuals from the protected group and that from the unprotected group are the same.

When considering discrimination from the perspective of supervised learning, the equality of opportunity is based on the actual outcome Y and the predicted outcome $\hat{Y}$, requiring $P(\hat{Y} = y^+|s^+, y^+) = P(\hat{Y} = y^+|s^-, y^+)$. Basically it means the decision model should not mistakenly predict examples with y^+ as $\hat{Y} = y^-$ at a higher rate for one group than another. In other words, a predictor $\hat{Y}$ satisfies

Table 1: Formula of previous fairness notions.

Notation	References	Formula
Demographic parity	[24]	$P(y^+\|s^+) - P(y^+\|s^-)$
Conditional parity	[24]	$P(y^+\|s^+, \mathbf{o}) - P(y^+\|s^-, \mathbf{o})$
Total causal discrimination	[29, 34]	$\mathbb{E}[Y(s^+)] - \mathbb{E}[Y(s^-)]$
Path-specific causal discrimination	[16, 34]	$\mathbb{E}[Y(s^+)\|\pi] - \mathbb{E}[Y(s^-)\|\pi]$
Counterfactual fairness	[11]	$\mathbb{E}[Y_{\mathbf{o}}(s^+)] - \mathbb{E}[Y_{\mathbf{o}}(s^-)]$
Path-specific counterfactual fairness	[26]	$\mathbb{E}[Y_{\mathbf{o}}(s^+)\|\pi] - \mathbb{E}[Y_{\mathbf{o}}(s^-)\|\pi]$
Equality of opportunity	[5, 27]	$P(\hat{Y} = y^+\|s^+, y^+) - P(\hat{Y} = y^+\|s^-, y^+)$
Calibration	[5, 27]	$P(y^+\|s^+, \hat{Y} = y^+) - P(y^+\|s^-, \hat{Y} = y^+)$

equalized opportunity with respect to protected attribute S and outcome Y if $\hat{Y}$ and S are independent conditional on Y. Similarly the calibration considers the fraction of correct positive predictions and requires $P(y^+|s^+, \hat{Y} = y^+) = P(y^+|s^-, \hat{Y} = y^+)$. Different from the previous methods that focuses on prediction results, our proposed equality of opportunity focuses on the effort, i.e., the minimum change of T to achieve a certain outcome level Y, based on the causal framework.

We noticed a parallel work [6] that developed an effort-based measure of fairness and formulated effort unfairness as the inequality in the amount of effort required for members from disadvantage group and advantaged group. However, their work focused on characterizing the long-term impact of algorithmic policies on reshaping the underlying population based on the psychological literature on social learning and the economic literature on equality of opportunity. Our work is based on counterfactual causal inference and develops an optimization-based framework for removing discriminatory effort unfairness from the static data if discrimination is detected.

4 CALCULATING AVERAGE EFFORT DISCREPANCY

In real-world applications, we often have multiple values of γ used in decision making. We use the average effort discrepancy over all values of γ as the measure of equality of effort in this scenario. If γ has a set of discrete values, then the average is computed by the mean of all effort discrepancies. If γ is a continuous variable, then the average is defined as the integration over the range of γ.

Definition 4.1 (Average Effort Discrepancy (AED)). If $\gamma \in \Gamma$ where Γ denotes the effort level value set of the expectation of outcome variable, then the average effort discrepancy is defined as

$$AED = \frac{1}{|\Gamma|} \sum_{\gamma \in \Gamma} \delta(\gamma), \tag{1}$$

where $\delta(\gamma)$ could be either $\delta_i(\gamma)$ or $\delta_D(\gamma)$.

If γ is a continuous variable in a range $[\gamma_1, \gamma_2]$, then the average effort discrepancy is defined as

$$AED = \frac{1}{\gamma_2 - \gamma_1} \int_{\gamma_1}^{\gamma_2} \delta(\gamma) d\gamma. \tag{2}$$

To calculate the AED, we need to first compute the expected outcome $\mathbb{E}[Y_{I^*}(t)]$ or $\mathbb{E}[Y_{D^*}(t)]$, and then compute the minimum effort. In the following, we develop a general calculating method assuming the monotonicity and invertibility for $\mathbb{E}[Y_{D^*}(t)]$. Then, we consider three widely used techniques for causal inference: outcome regression and propensity score weighting from Rubin's framework, and structural causal analysis from Pearl's framework. We compute the AED for each of the techniques.

Algorithm 1 Discrimination detection through equal effort.

Input: Dataset D, Threshold τ
Output: Discrimination detection result
1: For each subset $D^* \in \{D^+, D^-\}$, identify expected outcome $f_{D^*}(t) = \mathbb{E}[Y_{D^*}(t)]$
2: **if** $f_{D^*}(t)$ is continuous, monotonous and invertible **then**
3: Calculate *AED* according to Eq. (3)
4: **else**
5: Identify inverse function $f^{-1}_{D^*}(\gamma)$
6: **if** $f^{-1}_{D^*}(\gamma)$ has a closed form **then**
7: **for** each γ **do**
8: Find the minimum value of t such that $t \geq f^{-1}_{D^*}(\gamma)$
9: Calculate effort discrepancy $\delta_D(\gamma)$
10: **end for**
11: **else**
12: **for** each treatment level t **do**
13: Use appropriate causal inference method to estimate $\hat{\mathbb{E}}[Y_{D^*}(t)]$
14: **end for**
15: **for** each γ **do**
16: Numerically find the minimum value of t such that $\hat{\mathbb{E}}[Y_{D^*}(t)] \geq \gamma$
17: Calculate effort discrepancy $\delta_D(\gamma)$
18: **end for**
19: Calculate *AED* following Definition 4.1
20: **end if**
21: **end if**
22: **if** $|AED| \geq \tau$ **then**
23: Result = True
24: **else**
25: Result = False
26: **end if**

Algorithm 1 shows the pseudocode of our algorithm for computing the AED and making the judge of discrimination through equal effort. Lines 2-3 deal with the situation where $f_{D^*}(t) = \mathbb{E}[Y_{D^*}(t)]$ is a continuous, monotonous and invertible function of t, and AED can be directly computed through an integration over $f_{D^*}(t)$ given in the next subsection. If the assumptions are not satisfied, lines

6-10 handle the situation where the closed-form of inverse function $f^{-1}_{D^*}(\gamma)$ can be derived; and lines 12-19 handle the situation otherwise.

4.1 General Method under Monotonicity and Invertibility Assumption

As discussed in the previous section, $\mathbb{E}[Y_{D^+}(t)]$ and $\mathbb{E}[Y_{D^-}(t)]$ denote the expectations of outcome variable for groups D^+ and D^-. We can treat them as functions of t, denoted as $f_{D^+}(t)$ and $f_{D^-}(t)$. Under the assumptions of being monotonically increasing and invertible, inequality $\mathbb{E}[Y_{D^+}(t)] \geq \gamma$ can be expressed as $f_{D^+}(t) \geq \gamma$, which leads to $t \geq f^{-1}_{D^+}(\gamma)$, where $f^{-1}_{D^+}(\cdot)$ is the inverse function of $f_{D^+}(\cdot)$. As a result, we directly obtain that $\Psi_{D^+}(\gamma) = f^{-1}_{D^+}(\gamma)$, and similarly $\Psi_{D^-}(\gamma) = f^{-1}_{D^-}(\gamma)$.

If the closed forms of $f^{-1}_{D^+}(\cdot)$ and $f^{-1}_{D^-}(\cdot)$ can be derived, then the AED can be easily computed; otherwise its calculation is not straightforward. However, when γ is a continuous variable, then we don't need to derive the closed form of the inverse functions to compute the AED, but only require the integration of $f_{D^+}(\cdot)$ and $f_{D^-}(\cdot)$ to be tractable. This is because based on the Laisant's theorem we have

$$\int_{\gamma_1}^{\gamma_2} f^{-1}_{D^+}(\gamma)d\gamma = \gamma_2 t_2^+ - \gamma_1 t_1^+ - \int_{t_1^+}^{t_2^+} f_{D^+}(\gamma)d\gamma,$$

where $t_1^+ = f^{-1}_{D^+}(\gamma_1)$ and $t_2^+ = f^{-1}_{D^+}(\gamma_2)$. In practice, t_1^+ and t_2^+ can be estimated using numerical methods. As a result, the AED is given by

$$(t_2^+ - t_2^-)\gamma_2 - (t_1^+ - t_1^-)\gamma_1 - \left(\int_{t_1^+}^{t_2^+} f_{D^+}(\gamma)d\gamma - \int_{t_1^-}^{t_2^-} f_{D^-}(\gamma)d\gamma\right). \quad (3)$$

4.2 Outcome Regression

Outcome regression is one straightforward method to conduct causal inference. In this approach, a model is posited for the outcome variable as a function of the treatment variable and the covariates. The basic outcome regression model is the linear regression of the form:

$$\mathbb{E}[Y|T, \mathbf{X}] = \beta_0 + \beta_1 T + \boldsymbol{\beta_2 X} + \boldsymbol{\beta_3 X} T,$$

where β_0, β_1 are regression coefficients, $\boldsymbol{\beta_2}$ and $\boldsymbol{\beta_3}$ are the coefficient vectors with the same length as $\boldsymbol{X}$. All the parameters can be estimated by least squares method.

One advantage of outcome regression is it can help us directly calculate the relative treatment value given a certain expected outcome level. Suppose the regression model is correctly specified, the expected outcome of any subset D^* is given by

$$\mathbb{E}[Y_{D^*}(t)] = \frac{1}{|D^*|} \sum_{i \in D^*} (\beta_0 + \beta_1 t + \boldsymbol{\beta_2 x}_i + \boldsymbol{\beta_3 x}_i t).$$

Thus, the minimum value of the treatment variable to achieve γ-level outcome, i.e., $\Psi_{D^*}(\gamma)$, can be expressed as:

$$\underset{t \in T}{\operatorname{argmin}} \left\{\mathbb{E}[Y_{D^*}(t)] \geq \gamma\right) = \frac{\gamma - \frac{1}{|D^*|} \sum_{i \in D^*} (\beta_0 + \boldsymbol{\beta_2})}{\frac{1}{|D^*|} \sum_{i \in D^*} (\beta_1 + \boldsymbol{\beta_3})}. \quad (4)$$

4.3 Propensity Score Weighting

Another widely used branch of causal inference is based on weighting and one typical method is the inverse propensity score weighting. In our context, the treatment variable is a multiple valued ordinal variable, we apply generalized propensity score [8] to estimate the weights.

Definition 4.2 (Generalized Propensity Score). The generalized propensity score for individual i is the conditional probability of receiving a particular level of the treatment given the pre-treatment variables:

$$r(t, \mathbf{x_i}) = Pr(T = t|\mathbf{X_i} = \mathbf{x_i}).$$

The weighted mean of the potential outcomes for those who received the treatment t had they received another treatment t' can be consistently estimated by

$$\hat{\mathbb{E}}[Y(t')|t] = \frac{\sum_{i \in N} \mathbb{1}_{T_i = t'} Y_i \omega_i(t, t')}{\sum_{i \in N} \mathbb{1}_{T_i = t'} \omega_i(t, t')},$$

where

$$\omega_i(t, t') = \frac{r(t, \mathbf{x_i})}{r(t', \mathbf{x_i})}.$$

Following the above method, we can get a table showing estimation values of the expected outcome under all treatment pair combinations (t, t'). Thus, the minimum treatment value to achieve $\hat{\mathbb{E}}[Y(t')|t] \geq \gamma$ can be determined by comparing the results in that table.

4.4 Structural Causal Model

The structural causal model describes the causal mechanisms of a system as a set of structural equations. For ease of representation, each causal model can be illustrated by a directed acyclic graph called the causal graph, where each node represents a variable and each edge represents the direct causal relationship specified by the causal model. In addition, each node V is associated with a conditional probability distribution $P(v|\text{pa}_V)$ where pa_V is the realization of a set of variables PA_V called the parents of V. The treatment is modeled using the intervention, which forces the treatment variable T to take certain value t, formally denoted by $do(T = t)$ or $do(t)$. The potential outcome of variable Y under intervention $do(t)$ is denoted as Y_t. The distribution of Y_t, also referred to as the post-intervention distribution of Y under $do(t)$, is denoted as $P(Y_t)$. Facilitated by the intervention, the expected outcome $\mathbb{E}[Y_{D^*}(t)]$ can be measured by the counterfactual quantity $\mathbb{E}[Y_t|\mathbf{z}^*]$, where $\mathbf{z}^*$ represents attribute values that form the subgroup D^*. The counterfactual quantity measures the expected outcome of Y assuming that the intervention is performed on the subgroup of individuals only. According to [18], if attributes $\mathbf{Z}$ are non-descendant of T in the causal graph, then $P(Y_t|\mathbf{z}^*)$ can be computed from observational data as

$$\frac{\sum_{\mathbf{X} \setminus \mathbf{Z}} \prod_{V \in \{Y, S, \mathbf{X}\}} P(v|\text{pa}_V)_{\delta_{T=t}}}{P(\mathbf{z}^*)},$$

where $\delta_{T=t}$ means assigning T involved in all probabilities with the corresponding value t.

If the inverse function of $\mathbb{E}[Y_t|\mathbf{z}^*]$ can be derived, then we follow lines 6-10 in Algorithm 1 to compute AED; otherwise, we follow lines 12-19 to compute AED.

Table 2: Preprocessing *education*.

Category	Original Values
0	Preschool, 1st-4th, 5th-6th
1	7th-8th, 9th, 10th, 11th
2	12th, HS-grad, Some-college, Assoc-voc
3	Assoc-acdm, Bachelors, Masters, Prof-school
4	Doctorate

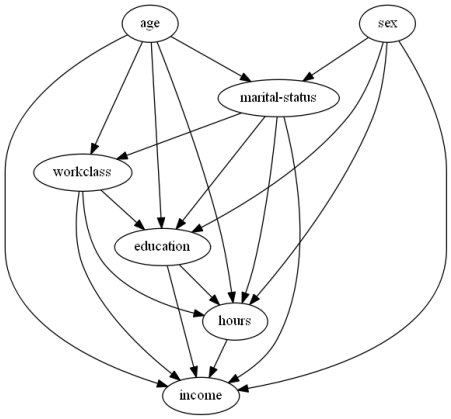

Figure 1: The causal graph for the Adult dataset.

5 ACHIEVING EQUAL EFFORT

When our discrimination detection algorithm shows that a dataset does not satisfy the equal effort requirement, then we may want to remove the discriminatory effects from the dataset before it is used for any predictive analysis, i.e., training a decision model. In this section, we develop a method for generating a new dataset which is close to the original dataset and also satisfies equal effort. Our removal method is based on the use of outcome regression to estimate the potential outcome, but it can be easily extended to any method where the closed form of $\Psi(\gamma)$ can be derived. The general idea is to derive a new outcome regression model satisfying the equal effort constraints. Then, for each individual in the original dataset, we randomly generate a new value $\tilde{Y}$ based on the expectation computed from the fair outcome regression model.

Specifically, we consider two outcome regression models for subsets D^+ and D^- respectively, given by

$$\mathbb{E}[Y_{D^+}|T,\mathbf{X}] = \beta_0^+ + \beta_1^+ T + \boldsymbol{\beta}_2^+ \boldsymbol{X} + \boldsymbol{\beta}_3^+ \boldsymbol{X} T,$$

$$\mathbb{E}[Y_{D^-}|T,\mathbf{X}] = \beta_0^- + \beta_1^- T + \boldsymbol{\beta}_2^- \boldsymbol{X} + \boldsymbol{\beta}_3^- \boldsymbol{X} T.$$

Then, as shown by Eq. (4), the minimum effort for subgroup D^+ (and similarly for subgroup D^-) is given by

$$\Psi_{D^+}(\gamma) = \frac{\gamma - \frac{1}{|D^+|}\sum_{i\in D^+}(\beta_0^+ + \boldsymbol{\beta}_2^+)}{\frac{1}{|D^+|}\sum_{i\in D^+}(\beta_1^+ + \boldsymbol{\beta}_3^+)}.$$

As a result, the AED according to either Eq. (1) or (2) is given by

$$\frac{\bar{\gamma} - \frac{1}{|D^+|}\sum_{i\in D^+}(\beta_0^+ + \boldsymbol{\beta}_2^+)}{\frac{1}{|D^+|}\sum_{i\in D^+}(\beta_1^+ + \boldsymbol{\beta}_3^+)} - \frac{\bar{\gamma} - \frac{1}{|D^-|}\sum_{i\in D^-}(\beta_0^- + \boldsymbol{\beta}_2^-)}{\frac{1}{|D^-|}\sum_{i\in D^-}(\beta_1^- + \boldsymbol{\beta}_3^-)},$$

where $\bar{\gamma}$ equals $\frac{1}{|\Gamma|}\sum_{\gamma\in\Gamma}\gamma$ if discrete and $\frac{\gamma_2^2-\gamma_1^2}{2}$ if continuous. We want the AED to approach zero. After adding the penalty term for the AED, the objective function becomes

$$\underset{\beta}{\operatorname{argmin}} \sum_{i\in D^+, D^-} (\mathbb{E}[Y_{D^*}|t_i,\mathbf{x}_i] - y_i)^2 + \lambda \cdot AED^2$$

where $D^* = D^+$ or D^- and λ is the parameter for balancing the two objectives.

Finally, for each individual i in the dataset with profile $(s_i, t_i, \mathbf{x}_i, y_i)$, we first compute his expected value of Y using the fair outcome regression model, i.e., $\mathbb{E}[Y_{D^*}|t_i,\mathbf{x}_i]$, where $D^* = D^+$ or D^- depending on the value of s_i. Then, we randomly assign 0 or 1 to the new value $\tilde{y}_i$ based on the probability given by $\mathbb{E}[Y_{D^*}|t_i,\mathbf{x}_i]$. The generated data then satisfies the equal effort requirement.

6 EXPERIMENTS

We evaluate our discrimination detection and removal algorithms based on the proposed equality of effort on the UCI Adult dataset [13]. The Adult dataset contains 65, 123 records with 14 attributes. We select 7 attributes, *sex*, *age*, *marital status*, *workclass*, *education*, *hours*, and *income* in our experiments. We consider *income* as the outcome, *education* as the treatment attribute, and *sex* as the protected attribute. Due to the sparse data issue, we binarize the domain of *age*, *marital status*, *workclass*, and *hours* into two classes. We also categorize 16 values of *education* into five levels, as shown in Table 2.

In our experiments, we calculate the minimum effort based on three methods, outcome regression (*Regression*), propensity score weighting (*Weighting*), and structural causal model inference (*SCM*). For *Weighting*, we implement the propensity score weighting for multiple treatments by following the work of [15] and [1]. For *SCM*, we follow the settings of [34] and use three tiers for causal graph learning: *sex*, *age* in Tier 1, *marital-status*, *education*, *workclass*, and *hours* in Tier 2, and *income* in Tier 3. The causal graph is constructed and presented by utilizing the open-source software TETRAD [22]. We employ the original PC algorithm [23] and set the significance threshold 0.01 for conditional independence setting in causal graph construction. Figure 1 shows the built causal graph. We apply the nonparametric inference of the structural causal model by following the work of [33]. In discrimination removal, the quadratic programming is solved using PyTorch [17].

6.1 Discrimination Discovery

6.1.1 Checking equal effort at the system level. Table 3 shows the comparison results of the expectations of the potential outcome for males ($\mathbb{E}[Y_{D^+}(t)]$) and that for females ($\mathbb{E}[Y_{D^-}(t)]$) in Adult. We calculate the expectation of the potential outcomes using three methods, *Weighting*, *Regression*, and *SCM*, and vary the treatment variable *education* from 0 to 4. As shown in Table 3, the expectations of potential outcome for males are significantly higher than the corresponding values for females, indicating large effort discrepancy exists in Adult. For example, $\mathbb{E}[Y_{D^+}(t)] = 0.498$ and $\mathbb{E}[Y_{D^-}(t)] = 0.221$ when $t = 2$ based on *SCM*. If we set $\gamma = 0.7$, the minimum values of treatment variable (*education*) to achieve γ-level outcome are 3 for males (with the expectation of the potential outcome 0.741) and 4 for females (with the expectation of the

Table 3: Expectation of the potential outcome of Attribute *income* for males and females in Adult dataset.

education	*sex=male*			*sex=female*		
	Weighting	*Regression*	*SCM*	*Weighting*	*Regression*	*SCM*
0	0.196	0.086	0.164	0.048	0.026	0.057
1	0.269	0.214	0.239	0.066	0.051	0.075
2	0.513	0.491	0.498	0.211	0.190	0.221
3	0.736	0.781	0.741	0.416	0.497	0.469
4	0.842	0.933	0.859	0.485	0.807	0.706

Table 4: Expectations of the potential outcome of Attribute *income* for males and females with the original *education*=0.

education	*sex=male*			*sex=female*		
	Weighting	*Regression*	*SCM*	*Weighting*	*Regression*	*SCM*
1	0.225	0.232	0.227	0.071	0.084	0.081
2	0.457	0.462	0.467	0.205	0.205	0.224
3	0.692	0.694	0.719	0.418	0.411	0.497
4	0.810	0.870	0.842	0.497	0.693	0.754

Table 5: Expectation of the potential outcome of Attribute *income* for three randomly chosen individuals.

education	*User 1*		*User 2*		*User 3*	
	sex=male	*sex=female*	*sex=male*	*sex=female*	*sex=male*	*sex=female*
0	-	-	-	-	0.012	0.006
1	0.022	0.007	0.058	0.030	0.051	0.024
2	0.085	0.036	0.206	0.134	0.188	0.096
3	0.282	0.159	0.523	0.438	0.501	0.317
4	0.624	0.487	0.823	0.796	0.813	0.669

potential outcome 0.706). The effort discrepancy between females and males is 1, which indicates the existence of significant discrimination in terms of equal effort fairness. We would like to point out that the expectations of potential outcome calculated from three methods are generally consistent as shown in Table 3. However, each calculation method has its own applicable assumptions and may not achieve reliable results when those assumptions are not met. There are extensive researches on the applicability of those causal inference methods (e.g., refer to [18]), which are out of the scope of this work.

6.1.2 Checking equal effort at the group level. For the group level equality of effort, we split the Adult dataset into five groups by *education*. Individuals with the same education value form one group. For each group, we calculate the expectations of potential outcome for males ($\mathbb{E}[Y_{D^+}(t)]$) and females ($\mathbb{E}[Y_{D^-}(t)]$). Due to space limit, we only report in Table 4 the expectations of the potential outcome variable for group one with *education=0*. Each expectation is calculated using three methods. We can see the significant discrepancy between males and females in this group. We also observe the similar phenomena in other four groups. When considering $\gamma = 0.5$, the minimum education value to achieve the outcome for males in this group is 3 (with all expectation values from three methods close to 0.7) whereas the minimum education level for females is 4.

6.1.3 Checking equal effort at the individual level. To detect effort discrepancy at the individual level, we need to first identify a subset of users I with the same characteristics of the given individual and then split them into the male group (I^+) and female group (I^-). We then calculate the expectations of potential outcome for the male group ($\mathbb{E}[Y_{I^+}(t)]$) and female group ($\mathbb{E}[Y_{I^+}(t)]$) with each treatment level t. Due to space limit, we only report in Table 5 the results of three randomly chosen female users whose index numbers are 425, 9569, and 46437. Both users 1 and 2 have the original education value 1 and user 3 has education value 0. As shown in Table 5, the expectations of outcome for I^+ are consistently higher than I^-, indicating the existence of discrimination in terms of equal effort for these three individuals. For example, results of user 3 show that the minimum effort for her to achieve 0.5-level outcome is education $t = 4$ whereas the corresponding minimum effort to achieve the same level outcome is $t = 3$ had she been a male.

6.2 Discrimination Removal

We run our removal algorithm to remove discrimination in terms of equality of effort from the Adult dataset, and then run the discovery algorithm to further examine whether discrimination is truly removed in the modified dataset. For comparison, we include the removal algorithm (Denoted by DI) of [4], which removes discrimination from the demographic parity perspective. Basically, DI tries to modify X such that the modified $\hat{X}$ cannot be used to predict S. The results show that, after executing our removal method (with $\lambda = 5$), the average difference between $\mathbb{E}[Y_{D^+}(t)]$ and $\mathbb{E}[Y_{D^-}(t)]$ for all ts is -0.0136, indicating all effort discrepancy has been removed. However, the average difference for the DI algorithm is 0.2628, showing that DI does not remove effort discrepancy. Regarding

data utility loss in terms of χ^2, our method also outperforms the DI algorithm in that the utility loss of our method is 34778, while the utility loss of the DI algorithm is 37997.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a new causality-based fairness notion called the equality of effort. Although previous notions can be used to judge discrimination from various perspectives (e.g., demographic parity, equal opportunity), they cannot quantify the (difference in) efforts that individuals need to make in order to achieve certain outcome levels. Our proposed notion, on the other hand, can help answer counterfactual questions like "how much credit score an applicant should improve such that the probability of her loan application approval is above a threshold", and judge discrimination from the equal-effort perspective. To quantify the average effort discrepancy, we developed a general method under certain assumptions and specific methods based on three common causal inference techniques. When equality of effort is not achieved in a dataset, we developed an optimization method to remove discrimination. In the experiments, we show that the Adult dataset does contain effort discrepancy at system, group, and also individual levels, and our removal method can ensure the newly generated dataset satisfies equality of effort.

We made several assumptions in our paper including the no-hidden-confounder assumption, monotonicity of the expectation of outcome variable, and invertibility of outcome function. We also assumed one binary protected attribute and one binary decision for simplicity's sake. The no-hidden-confounder assumption is a common assumption for causal inference [18] and widely adopted by causal inference based fair learning. The monotonicity assumption reflects the real world phenomena (the more effort, the better outcome). The invertibility assumption is used in our general method of calculating the average effort discrepancy without deriving the closed form of the inverse function. When this invertibility assumption is not held, we have presented in our algorithm (Lines 12-19) several inference methods that could also have their limitations. Moreover, we implicitly assumed that the discrimination detection algorithm knows the same information as the decision-maker, i.e., there are no omitted variables used in decision making but invisible to the discrimination detection. In our future work, we will study how to achieve equal effort fairness when some of those assumptions are not met in practice.

In our paper, we used the change of treatment variable value as the effort needed to achieve a certain level of outcome and did not consider the real monetary or resource cost behind that change that are often not included in the data. If they are included in the data, the discrimination caused by these factors is known as indirect discrimination. We will study the use of path-specific effect/mediator analysis [16, 34] to explicitly quantify the effect of treatment on final outcomes via proxy attributes.

ACKNOWLEDGMENTS

This work was supported in part by NSF 1646654, 1920920, 1937010, and 1940093

REFERENCES

[1] Lane Burgette, Beth Ann Griffin, and Dan McCaffrey. 2017. Propensity scores for multiple treatments: A tutorial for the mnps function in the twang package. *R package. Rand Corporation* (2017).

[2] Silvia Chiappa. 2019. Path-Specific Counterfactual Fairness. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, the Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, the Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. 7801–7808. https://doi.org/10.1609/aaai.v33i01.33017801

[3] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM, 214–226.

[4] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*. ACM Press, 259–268. https://doi.org/10.1145/2783258.2783311

[5] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[6] Hoda Heidari, Vedant Nanda, and Krishna Gummadi. 2019. On the Long-Term Impact of Algorithmic Decision Policies: Effort Unfairness and Feature Segregation through Social Learning. In *International Conference on Machine Learning*. 2692–2701.

[7] Paul W Holland. 1986. Statistics and causal inference. *Journal of the American statistical Association* 81, 396 (1986), 945–960.

[8] Guido W Imbens. 2000. The role of the propensity score in estimating dose-response functions. *Biometrika* 87, 3 (2000), 706–710.

[9] Aria Khademi, Sanghack Lee, David Foley, and Vasant Honavar. 2019. Fairness in Algorithmic Decision Making: An Excursion Through the Lens of Causality. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*. 2907–2914. https://doi.org/10.1145/3308558.3313559

[10] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*. 656–666.

[11] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.

[12] Jiuyong Li, Jixue Liu, Lin Liu, Thuc Duy Le, Saisai Ma, and Yizhao Han. 2017. Discrimination detection by causal effect estimation. In *2017 IEEE International Conference on Big Data, BigData 2017, Boston, MA, USA, December 11-14, 2017*. IEEE, 1087–1094. https://doi.org/10.1109/BigData.2017.8258033

[13] M Lichman. 2013. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml.

[14] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 2011. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 502–510.

[15] Daniel F McCaffrey, Beth Ann Griffin, Daniel Almirall, Mary Ellen Slaughter, Rajeev Ramchand, and Lane F Burgette. 2013. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in medicine* 32, 19 (2013), 3388–3414.

[16] Razieh Nabi and Ilya Shpitser. 2018. Fair Inference on Outcomes. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. 1931–1940.

[17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).

[18] Judea Pearl. 2009. *Causality*. Cambridge university press.

[19] Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* 29, 05 (2014), 582–638.

[20] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.

[21] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. 2017. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*. 6414–6423.

[22] Richard Scheines, Peter Spirtes, Clark Glymour, Christopher Meek, and Thomas Richardson. 1998. The TETRAD project: Constraint based aids to causal model specification. *Multivariate Behavioral Research* 33, 1 (1998), 65–117.

[23] Peter Spirtes, Clark N Glymour, and Richard Scheines. 2000. *Causation, prediction, and search*. Vol. 81. MIT press.

[24] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 1–7.

[25] Yongkai Wu, Lu Zhang, and Xintao Wu. 2019. Counterfactual Fairness: Unidentification, Bound and Algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. International Joint Conferences on Artificial Intelligence Organization, 1438–1444. https://doi.org/10.24963/ijcai.2019/199

[26] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. 2019. PC-Fairness: A Unified Framework for Measuring Causality-Based Fairness. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, December 8-14, 2019, Vancouver, Canada, 2019*. Curran Associates, Inc., 3399–3409.

[27] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS*.

[28] Junzhe Zhang and Elias Bareinboim. 2018. Equality of Opportunity in Classification: A Causal Approach. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*. 3675–3685.

[29] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in Decision-Making - the Causal Explanation Formula. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. 2037–2045.

[30] Lu Zhang and Xintao Wu. 2017. Anti-discrimination learning: a causal modeling-based framework. *International Journal of Data Science and Analytics* 4, 1 (May 2017), 1–16. https://doi.org/10.1007/s41060-017-0058-x

[31] Lu Zhang, Yongkai Wu, and Xintao Wu. 2016. On Discrimination Discovery Using Causal Networks. In *Social, Cultural, and Behavioral Modeling, 9th International Conference, SBP-BRiMS 2016, Washington, DC, USA, June 28 - July 1, 2016, Proceedings*, Vol. 9708. Springer, 83–93. https://doi.org/10.1007/978-3-319-39931-7_9

[32] Lu Zhang, Yongkai Wu, and Xintao Wu. 2016. Situation Testing-Based Discrimination Discovery: A Causal Inference Approach. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, Vol. 2016-Janua. IJCAI/AAAI Press, 2718–2724.

[33] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. Achieving Non-Discrimination in Data Release. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*. ACM Press, New York, New York, USA, 1335–1344.

[34] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017*. 3929–3935.

[35] Lu Zhang, Yongkai Wu, and Xintao Wu. 2019. Causal modeling-based discrimination discovery and removal: criteria, bounds, and algorithms. *IEEE Trans. Knowl. Data Eng.* 31, 11 (2019), 2035–2050. https://doi.org/10.1109/TKDE.2018.2872988

[36] Indre Žliobaite, Faisal Kamiran, and Toon Calders. 2011. Handling conditional discrimination. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. IEEE, 992–1001.