The Illusion of Fairness: Auditing Fairness Interventions in Algorithmic Hiring with Audit Studies

Disa Sariola[1], **Patrick Button**[1, 2], **Aron Culotta**[1], **Nicholas Mattei**[1]
[1]Tulane University, New Orleans, LA, USA
[2]Connolly Alexander Institute for Data Science, New Orleans, LA, USA
dsariola@tulane.edu, pbutton@tulane.edu, aculotta@tulane.com, nsmattei@tulane.edu

Abstract

Classifiers trained on historical data are deployed in the real world to automate decisions from hiring to loan issuance. Judging the fairness and efficiency of these systems, and their human counterparts, is a complex and important topic studied across both computational and social sciences. One common way to address bias in classifiers is to resample the training data to offset distributional disparities. In the hiring domain, where results may vary by a protected class, many interventions from the literature equalize the hiring rate within the training set to alleviate bias. While simple and seemingly effective, these methods have typically only been evaluated using data obtained through convenience samples, e.g., data from a real-world hiring process, introducing selection and label bias. In the social and health sciences, audit studies, in which fictitious "testers" (resumes) are sent to subjects (job openings) in a randomized control trial, provide high-quality data that support rigorous estimates of discrimination by controlling for confounding factors. We investigate how data from audit studies can be used to improve our ability to both train and evaluate automated hiring algorithms. Specifically, we use data from a large audit study of age discrimination in hiring to test common resampling methods from the fair machine learning literature. We find that audit data of real-world hiring reveals cases where equalizing base rates across classes *appears* to achieve parity using traditional measures, but in fact has an absolute $\approx 10\%$ disparity when measured appropriately. We also show that corrections based on individual treatment effect estimation methods combined with audit study data can overcome these issues, underscoring the need for rigorous data collection in fairness research.

Code —
https://github.com/AlexandraSar/IllusionOfFairness
Extended version — https://arxiv.org/abs/2507.02152

1 Introduction

A foundational assumption in most work on algorithmic discrimination is that the human-provided class labels in the training data are improperly influenced, directly or indirectly, by a protected or minoritized characteristic like age, race, gender, or sexual orientation. Despite this assumption, we often evaluate machine learning methods on the very same flawed data, leading to biased estimates of both fairness and accuracy (Mehrabi et al. 2021). While a truly unbiased annotation is infeasible for most tasks (e.g., recidivism prediction, hiring, loans), we consider how this issue can be addressed using data from audit studies of hiring.

In an audit study,[1] "testers" (e.g., resumes, inquiries, simulated patients) are randomly assigned covariates and protected characteristics and given to humans for assessment (e.g., callback interview, response, diagnosis). This randomization allows rigorous estimation of the amount of discrimination in a real-world system (Gaddis 2018). To find no discrimination, the difference in outcome (class) that can be attributed to the protected/minoritized variable should have no difference (on average) between the groups.

Audit studies ensure the only difference between testers is the protected attribute (e.g., age) by creating virtual twins for every applicant. Hence, any difference in outcomes is label bias. Traditional fairness metrics inherit that bias, while counterfactual based methods are able to remove it. While many deployed models are designed to ensure certain fairness metrics, they often ignore the human decisions within the data, treating observed labels as ground truth, not accounting for label bias. Audit studies measure discrimination by holding all attributes constant, varying only the protected ones. This eliminates hidden confounders and allow us to quantify the causal effect of a protected attribute.

We analyze data collected from a real-world audit study of hiring to understand age discrimination in applicant callback decisions. Specifically, 40,208 resumes were sent to 13,371 job openings across 11 cities in the United States (Neumark, Burn, and Button 2019), and information was recorded on which resumes received a callback to interview. The resumes were systematically constructed to isolate the effect of age on callback, controlling for other factors. Causal effects within hiring are a frequent topic of study for fairness in ML (Pearl 2010; Bogen and Rieke 2018).

Such a large and rigorously collected sample of human decisions serves as a useful testbed to investigate the behavior of machine learning systems trained to automate hiring decisions. In this paper, we study how such data may be

[1]Audit studies are also known as audit field experiments, correspondence studies, resume studies/experiments, simulated patient studies, and vignettes – henceforth "audit studies" for short.

used to improve both the training and evaluation of classifiers used in hiring and beyond. First, we explore how audit study data, unlike typical training data, enables robust estimates of the amount of *label bias* in the data — i.e., the quantity of labels (human decisions/outcomes) that should be changed to eliminate discrimination. Second, we evaluate an approach based on the popular notion of individual treatment effect (ITE) estimation to quantify the likelihood that each individual record has been subject to discrimination (i.e., has a biased label) (Corbett-Davies et al. 2023). Based on these estimates, we propose an algorithm to generate de-biased versions of the original data. We find this approach allows us both to better estimate the true accuracy and fairness of classifiers, as well as to improve the quality of classifiers trained on such data.

Our key contribution is a rigorous empirical analysis using audit study data in the hiring domain to evaluate:

RQ1: How does removing label bias with ITE *in the testing data* influence estimates of accuracy and fairness of classifiers trained on audit study data? We find that traditional approaches can create an *illusion of fairness*, in which methods that appear fair when evaluated using standard approaches are in fact shown to exhibit significant discrimination once label bias is reduced.

RQ2: How does removing label bias with ITE *in the training data* influence the accuracy and fairness of classifiers trained on audit study data? Training on debiased data reduces measures of disparity by up to 60% compared to traditional pre-processing approaches that equalize the base rate of protected attributes.

RQ3: How does the magnitude of discrimination in the human audit data influence the results? When we resample data to double the amount of human discrimination in the audit data, we find a commensurate increase in the inaccuracies of traditional fairness metrics. Our proposed ITE adjustment appears more robust, though classification accuracy does degrade in this setting.

RQ4: If audit study data is unavailable, how does selection bias affect the accuracy and fairness of resulting classifiers? When we reintroduce selection bias into the dataset, we find that the illusion of fairness can become more extreme, with measures of fairness diverging in magnitude and sign — e.g., a classifier that appears discriminatory against younger applicants is in fact discriminatory against older applicants.

2 Related Work

Auditing Systems and Processes Audit studies are field experiments where traits of real or hypothetical individuals are randomized to test their impact on outcomes (Gaddis 2018). "Testers" (e.g., resumes, emails) differ only in the traits being studied, on average, isolating the effects of bias. Correspondence studies are audit studies conducted remotely, e.g., through emails, messaging, or job applications. (Collins et al. 2021; Evans et al. 2015; Steiner, Atzmüller, and Su 2016). Recently, these approaches have also been applied to audit *algorithms* (Bandy 2021; Vecchione, Levy, and Barocas 2021), evaluating both their fairness and accuracy.

However, bias can originate from many directions – the model, the metrics, or the data (Hutchinson and Mitchell 2019). A central goal of "fair machine learning" systems is to prevent such harm across groups (Li, Goel, and Ash 2022a). This means that the result for subgroups should be comparable – people of similar circumstances should receive similar outcomes independent of protected characteristics. The larger framework of auditing and evaluating algorithms is a broad and active topic of research (Mitchell et al. 2021; Mehrabi et al. 2021).

Auditing Hiring The hiring process generally consists of three stages (Stredwick 2005; Bogen and Rieke 2018): job planning/analysis, opening advertisement/recruitment, and interview/hire. Like nearly all audit studies, we focus on the critical interview (callback) stage (Gaddis 2018).

Extensive research documents bias in traditional hiring (Lippens, Vermeiren, and Baert 2023). Audit studies show that older applicants, particularly women, receive fewer callbacks (Lahey 2008; Farber, Herbst, and Silverman 2019; Neumark, Burn, and Button 2019; Burn et al. 2020). Bias can be unintentional – e.g., when too many qualified resumes overwhelm reviewers or when irrelevant criteria are used, knowingly or not (Derous and Ryan 2018). Even minor instances of subgroup bias can lead to significant discrimination (Hardy et al. 2022). To reduce manual labor, some companies now use AI for resume screening, but this raises concerns over defining fair criteria. While anti-discrimination laws exist, ensuring fairness becomes harder with complex systems. Public perception of hiring algorithms is mixed, with many viewing them as less fair (Langer, König, and Papathanasiou 2019; Langer et al. 2020; Newman, Fast, and Harmon 2020).

Label Bias and Algorithmic Fairness Within algorithmic fairness (Barocas, Hardt, and Narayanan 2023), there is an emerging line of work focusing on the issue of *label bias* – a recognition that often the human decisions, which serve as ground truth, are themselves influenced by bias. Fish, Kun, and Lelkes (2016) were among the earliest to note this problem in algorithmic fairness, which they explored via simulation studies. Building on this, Wick, Panda, and Tristan (2019) propose that when biases such as selection and label bias are accounted for, the trade-off between accuracy and fairness can diminish or even disappear. To support this claim, they propose evaluating fairness when unbiased ground truth labels are available. As in Fish, Kun, and Lelkes (2016), simulation studies are required to explore data that does not have selection or label bias.

Similarly, Verma, Ernst, and Just (2021) focus on identifying and removing training instances affected by label bias in historical datasets. To identify such instances, they find matched pairs of instances that receive different labels, and assume that the one with the least confident classification decision is the one that has received discrimination. As above, this work relies on synthetically generated instances for training and testing. Finally, Jiang and Nachum (2020) present a mathematical framework for mitigating label bias by assuming that there is an existing underlying unbiased la-

bel function. They introduce a re-weighting scheme that adjusts the significance of some training instances to account for label bias. However, this approach is generally designed for fairness metrics that do not require unbiased labels (e.g., demographic parity (Dwork et al. 2012)). For other measures, additional assumptions and estimates are required.

Contribution. **(1)** Using audit study data to analyze algorithmic fairness, we are able to control for selection bias and rigorously quantify the amount of label bias present *without relying on synthetic data*. **(2)** We introduce a new method to correct for label bias when audit data is available using individual treatment effect (ITE) estimators. And **(3)** We provide empirical evidence demonstrating that traditional fairness evaluation metrics, when applied to conventionally biased labels that arise from samples of convenience, may produce misleading conclusions about algorithmic fairness.

3 Training and Evaluating Classifiers Using Human Audit Study Data

Our data is from a large-scale field experiment investigating age discrimination in hiring (Neumark, Burn, and Button 2019). Over 40,000 synthetic resumes were created and sent in response to online postings for four types of job positions: retail sales, administrative assistants, janitors, and security. The measure was whether the synthetic applicant received a callback from the potential employer. The study's aim was to determine whether callback rate was influenced by age, all else being equal, comparing young (age 29-31), middle-aged (age 49-51), and old (age 64-66) applicants.

The goal of such human audit studies is to rigorously estimate potential discrimination, considerable care is given to how resumes are created. Resumes of comparable skill and experience were sent to the same ad – varying only the age of the applicant – to isolate the effect of age on callback.[2] The study found strong overall evidence of age discrimination, with callback rates significantly lower for middle-aged ($\downarrow 18\%$) and older ($\downarrow 35\%$) applicants, as compared to younger applicants, despite the comparable resumes.

3.1 Training Classifiers on Audit Study Data

We study the behavior of a machine learning system trained to replicate the human decisions in the data. This simulates a scenario in which a firm attempts to automate the callback decision-making process based on historical decisions. As Neumark, Burn, and Button (2019) rigorously show, there is considerable discrimination present in this data; we wish to study its effect on the fairness and accuracy of resulting classifiers, as well as our ability to measure these values.

We train classifiers to predict the callback variable $Y \in \{0, 1\}$ given applicant attributes $\mathbf{X} \in \mathbb{R}^d$, and the protected age attribute $A \in \{y, o\}$, with labels y for younger applicants and o for older applicants.[3] Applicant attributes include demographics (gender, location), employment status, foreign language skills, typing speed, college experience, and volunteering history, see the extended version for more details on the data. One key variable we will explore is Spanish, which has a consistent positive correlation with callback in the audit data. We explore its use as a confounder since by design it is uniformly distributed in the audit data, independent of age. While there are studies on the sensitivity of unmeasured and unobserved confounders, in our case we know the effect of our confounders. Other studies of these phenomena often rely on simulations or parameterized models of these interactions that are not *directly present in the dataset, a key advantage of audit studies, and a fact that we are among the first to explore* (Kilbertus et al. 2020; Byun et al. 2024).

Age Group	Callback	No Callback	Total
Young	2,505 (19%)	10,896 (81%)	13,401
Old/Middle	3,587 (14%)	21,945 (86%)	25,532
Total	6,095 (16%)	32,892 (84%)	38,987

Table 1: Callback data by age group.

We use the audit study data as a labeled dataset $D = \{(\mathbf{X}_1, A_1, Y_1), \ldots, (\mathbf{X}_n, A_n, Y_n)\}$. Table 1 displays the callback rates by age group, showing a roughly 5% percentage point greater callback rate for younger applicants over older applicants. We experiment with two standard classification models, random forests and neural networks (Pedregosa et al. 2011), performing cross-validation to evaluate fairness and accuracy. As is typical in many studies on fairness, we focus on these standard models and not more complex ones as our goal is to understand the effects of the data itself (Machado, Charpentier, and Gallic 2025; Fawkes et al. 2024).

3.2 Evaluation Measures

To measure accuracy while accounting for class imbalance, we use Area Under the ROC Curve (**AUC**) (Fan, Upadhye, and Worster 2006). To measure fairness, we focus primarily on False Positive Rate Disparity (**FPRD**); FPRD is the difference in false positive classification rates between young and old applicants:

$$FPRD = FPR_y - FPR_o$$

with *false positive rate* defined as standard (Corbett-Davies et al. 2023):

$$FPR_y = \frac{FP_y}{FP_y + TN_y}, \quad FPR_o = \frac{FP_o}{FP_o + TN_o}$$

where FP (*false positives*) is the number of negative instances incorrectly classified as positive, TN (*true negatives*) is the number of negative instances correctly classified as negative, and $FP+TN$ is the total number of *actual negative* instances (AN). Additionally, we stratify across features, with a 20/80 split for test-train.

FPRD will be *positive* when the classifier discriminates against older applicants, *negative* when it discriminates against younger applicants, and *zero* when the classifier does not discriminate based on age. Within the hiring setting, a

[2]There are many nuances here (e.g., older applicants should have longer histories) (see Neumark, Burn, and Button (2019))

[3]For simplicity, we collapse middle and older applicants into a single age group called "older", with the remainder as "younger".

Group	AN	FP	FPR	FPRD
Y (Before)	100	30	0.3	**0.1**
O (Before)	100	20	0.2	
Y (↓ callbacks)	120	45	0.375	**0.3125**
O (↑ callbacks)	80	5	0.0625	

Table 2: Reducing label bias – increasing true callbacks (reducing actual negatives) for older applicants and decreasing true callbacks (increasing actual negatives) for younger applicants – affects False Positive Rate Disparity (FPRD).

false positive can be viewed as an applicant receiving a callback when they should not have. Thus, if FPRD is positive, then the rate at which unqualified younger applicants receive callbacks is higher than that of unqualified older applicants.

To ensure reliable comparison across classifiers with different positive prediction rates, we standardize the number of predicted positives by fixing a common callback budget. As the original dataset has an overall callback rate of 16%, we enforce the same rate during evaluation. For each classifier, we sort test instances by the predicted probability of a callback and label the top 16% as predicted positive.

4 Label Bias and the Illusion of Fairness

The primary problem in using human-provided labels from historical data, i.e., samples of convenience, to train and evaluate a classifier is the presence of *label bias* (Jiang and Nachum 2020). That is, for each applicant i, we only observe the label y_i, which we know is the result of a decision-making process influenced by age discrimination. Unfortunately, we cannot observe the idealized label y_i^* that would result from a process free of discrimination. This label bias will corrupt our measures of both accuracy and fairness. A classifier that prefers younger applicants to older applicants may appear more accurate, as it will better reflect human-generated labels. Likewise, measures of discrimination such as FPRD may be underestimated when computed using data with label bias. This is because removing label bias results in fewer young applicants getting callbacks and more older applicants getting callbacks, altering the false positive rates of each group. The audit study allows us to interrogate this in direct ways since across the dataset, applicants are constructed to be independent of confounders like age.

Consider the example in Table 2 of how removing label bias can reduce fairness. Initially, both groups have the same number of actual negatives (AN), though the classifier has a 10 percentage point higher FPR for younger. To remove label bias, we assume that 20 younger applicants who were given callbacks in the human audit study should not have been; of these, 15 were labeled as callbacks by the classifier. As a result, for the younger group AN_y increases to 120, while FP_y increases to 45, resulting in an increase in FPR_y from $0.3 \rightarrow 0.375$. Conversely, for the older group AN_o decreases to 80, while FP_o decreases to 5, resulting in a decrease in FPR_o from $0.2 \rightarrow 0.0625$. Hence, removing label bias affecting 40 applicants increased the FPRD estimate from $0.1 \rightarrow 0.3125$.

Algorithm 1: Repairing label bias with ITE

1: Fit a random forest classifier on training data D_{train}
2: Compute age ITE estimates $\hat{\tau}(\mathbf{x}_i)$ for each instance in the test set D_{test}
3: **while** callback rate is not equal between age groups **do**
4: # Flip positive to negative for younger group
5: Find i with largest $\hat{\tau}(\mathbf{x}_i)$ where $Y_i = 1$, $A_i = 1$
6: Set $Y_i \leftarrow 0$
7: # Flip negative to positive for older group
8: Find j with smallest $\hat{\tau}(\mathbf{x}_j)$ where $Y_j = 0$, $A_j = 0$
9: Set $Y_j \leftarrow 1$
10: **end while**

This illustrates how sensitive fairness metrics are to the presence of label bias, and underscores the need to more rigorously assess discrimination in labeled data. Otherwise, label bias can create an illusion of fairness, causing classifiers to *appear* less discriminatory than they are.

5 Repairing Label Bias with Individual Treatment Effect Estimation

If we were able to remove label bias, we would not only improve the reliability of our fairness measures, but also create cleaner training data to fit the classifier in the first place. However, doing so raises two questions: (1) Which applicants should have their callback labels amended? and (2) How many labels do we need to amend?

Typical samples of convenience make these questions difficult to answer. For example, if younger applicants tend to be more qualified in the applicant pool, it is not clear what the true callback rates should be for each group. In an audit study, however, by design we expect equal callback rates between the two age groups. This is the motivation for carefully controlling other resume attributes when constructing the synthetic resumes. Thus, the answer to our second question is: amend the labeled data until the callback rates are equal for younger and older applicants.

We amend labels in cases where age was a decisive factor in the callback decision, since those applicants are the ones whose outcomes were directly influenced by age. So, for younger applicants who received a callback, if they had been older applicants, would they still have received a callback? For older applicants who did not receive a callback, if they had been younger, would they have received a callback?

Framed this way, we have a counterfactual question – what would the outcome have been for an applicant if they had been in a different age group? To answer this, we build on a long line of work in the computational, social, and medical sciences on *individual treatment effect estimation* (Carey and Wu 2022; Plečko, Bareinboim et al. 2024).

Given the audit study data $D = \{(\mathbf{X}_1, A_1, Y_1), \ldots, (\mathbf{X}_n, A_n, Y_n)\}$, we treat the age variable A_i as a binary *treatment* indicator representing whether i is in the younger treatment ($A_i = 1$) group or older control ($A_i = 0$) group, and Y_i is the observed callback *outcome* for individual i. We are interested in quantifying the causal effect that treatment A has on the outcome Y. The fundamental problem of causal inference is

that we can only observe one outcome per individual, either the outcome of an individual receiving a treatment or not. Thus, we do not have direct evidence of what might have happened had we given individual i a different treatment.

Rubin's potential outcome framework is a common way to formalize this problem (Rubin 1974). Let $Y^{(1)}$ indicate the potential outcome an individual would have gotten had they received treatment ($A = 1$), and similarly let $Y^{(0)}$ indicate the outcome an individual would have gotten had they received no treatment ($A = 0$). While we cannot observe both $Y^{(1)}$ and $Y^{(0)}$ for the same individual, we can now formally express the quantity of interest. We are interested in the *Individual Treatment Effect* (ITE), which is the expected difference in outcome for a specific type of individual:

$$\tau(\mathbf{x}) = \mathbb{E}[Y^{(1)}|\mathbf{X} = \mathbf{x}] - \mathbb{E}[Y^{(0)}|\mathbf{X} = \mathbf{x}] \quad (1)$$

that is, the treatment effect for individuals where $\mathbf{X} = \mathbf{x}$. For example, if the covariate vector represents the (gender, height) of a person, then the ITE will estimate treatment effects for individuals that match along those variables.

Using standard assumptions, we estimate ITE as follows:

$$\hat{\tau}(\mathbf{x}) = \mathbb{E}[Y|A = 1, \mathbf{X} = \mathbf{x}] - \mathbb{E}[Y|A = 0, \mathbf{X} = \mathbf{x}] \quad (2)$$

$$= \frac{1}{|S_1(\mathbf{x})|} \sum_{i \in S_1(\mathbf{x})} Y_i - \frac{1}{|S_0(\mathbf{x})|} \sum_{i \in S_0(\mathbf{x})} Y_i \quad (3)$$

where $S_1(\mathbf{x})$ is the set of individuals i such that $\mathbf{X}_i = \mathbf{x}$ and $A_i = 1$, and similarly for $S_0(\mathbf{x})$. In other words, Equation (3) simply computes, for all individuals with covariates equal to $\mathbf{x}$, the difference between the average outcome for individuals in the treatment group and the control group. For example, if $\mathbf{X} = \mathbf{x}$ indicates individuals with (gender=male, height=5), $A = 1$ indicates that an individual is in the younger group, and $A = 0$ that they are in the older group, then $\hat{\tau}(\mathrm{x})$ is the difference in average outcome between individuals who are in the young group and those who are not.

A key challenge for Equation (3) in practice is that $\mathbf{X}$ may be high dimensional, leading to a small sample where $\mathbf{X} = \mathbf{x}$. In the extreme case, there may be only one instance where $\mathbf{X} = \mathbf{x}$. We adopt the virtual twins approach (Foster, Taylor, and Ruberg 2011), a two-step procedure to estimate ITE. First, it fits a random forest with all data (control and treatment samples), where each is represented by inputs $(\mathbf{X}_i, A_i)$ and outcome Y_i. Then, to estimate the ITE for i, it computes the difference between the predicted values for treatment $(\mathbf{X}_i, A_i = 1)$ and control $(\mathbf{X}_i, A_i = 0)$. The name "virtual twin" derives from the fact that for each control input $(\mathbf{X}_i, A_i = 0)$, we make a copy $(\mathbf{X}_i, A_i = 1)$ as treatment input that is alike in every way to the control input except for the treatment variable. Similarly, for each treatment input $(\mathbf{X}_i, A_i = 1)$, we make a copy $(\mathbf{X}_i, A_i = 0)$.

If $\hat{Y}(\mathbf{x}_i, 1)$ is the posterior probability of callback ($Y_i = 1$) from the random forest for input $(\mathbf{X}_i = \mathbf{x}_i, A_i = 1)$, and $\hat{Y}(\mathbf{x}_i, 0)$ is the probability of callback for input $(\mathbf{X}_i = \mathbf{x}_i, A_i = 0)$, then the virtual twin ITE estimate for i is

$$\hat{\tau}(\mathbf{x}_i) = \hat{Y}(\mathbf{x}_i, 1) - \hat{Y}(\mathbf{x}_i, 0). \quad (4)$$

Thus, $\hat{\tau}(\mathbf{x}_i)$ is the increase in probability of callback for being younger.

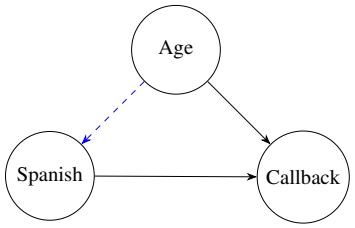


Figure 1: Causal diagram depicting our approach to injecting selection bias in audit study data. Note that because we use audit data there is no actual relationship between Spanish and Age in our dataset, but because we have audit data we can *simulate* such a relationship and observe the effects.

Algorithm 1 first computes all ITE estimates, then iteratively changes the callback labels most likely influenced by discrimination. At each iteration, we assign one younger applicant with a callback to have a no-callback label, and one older applicant with a no-callback label to have a callback label. We repeat until callback rates are equal between age groups, as expected in a world with no discrimination.

6 Experiments

The goal of our experiments is to answer the research questions from the introduction, comprehensively evaluating the effects of label bias on classifier training.

- RQ1: How does removing label bias with ITE *in the testing data* influence estimates of accuracy and fairness of classifiers trained on audit study data?
- RQ2: How does removing label bias with ITE *in the training data* influence the accuracy and fairness of classifiers trained on audit study data?
- RQ3: How does the magnitude of discrimination in the human audit data influence the results?
- RQ4: If audit study data is unavailable, how does selection bias affect the accuracy and fairness of resulting classifiers?

To do this, we consider three types of data pre-processing: We consider three types of data pre-processing:

1. **Base Rate (BR)**: Do not perform any manipulation of either the training or testing data.
2. **Equalized Base Rate (EBR)**: A simple yet common preprocessing technique to improve fairness is to downsample data to ensure equal class distributions across protected classes (Kleinberg, Mullainathan, and Raghavan 2016; Li, Goel, and Ash 2022b). To do so, we delete from the training set at random older applicants who did not get a callback until the callback rates are equal for the two age groups.
3. **Individual Treatment Effect Repair (ITE)**: The method from Algorithm 1.

To further investigate the impacts of these interventions separately on the training and testing data, we consider the following settings:

- **ITE Train & Test:** Apply ITE to both training and test. To do so, we perform cross-validation, estimating ITE for the test set in each fold. Then, for each fold, we apply Algorithm 1 separately for the train and test sets, ensuring equal callback rates between groups in both sets.
- **EBR Train - ITE Test:** Apply EBR to the training data, but ITE to the testing data.

Comparing EBR with (EBR Train - ITE Test) allows us to isolate the effect of label bias in the testing data on fairness.

We consider two classification models: **Random Forest** and **Multi-Layer Perceptron**. For RF, key parameters are: number of estimators=50, minimum samples per split=2, minimum samples per leaf=1. For MLP, we use three hidden layers of sizes (128, 64, 32), ReLU activation functions, and the Adam optimizer. All experiments are conducted in scikit-learn and available on GitHub (Pedregosa et al. 2011). Observe that we use AUC as our metric in the following, however the accuracy of the RF model is 77.62% (S.D. 0.0025) and MLP 78.32% (S.D. 0.0046). These results are in line with classifiers in many recent fairness papers (Zafar et al. 2017), indicating that our models are reasonable.

6.1 Simulating Selection Bias

For RQ4, we need data that reflects non-audit study data, i.e., we must reintroduce the sample selection bias that pervades samples of convenience normally used. Selection bias occurs when the distribution of data inadvertently introduces undesired correlations between the features pertaining to a protected attribute and the class label. Suppose that Spanish is predictive of callbacks, and that younger applicants are more likely to speak Spanish, illustrated below. Figure 1 displays the causal model for such a hypothetical scenario. The selection bias of Spanish may introduce an unintended relationship between age and callback. By contrast, *because our data is from an audit study, Spanish is not correlated with age by design*. However, we use Spanish as it has consistent positive correlation with callback, simulating this effect when we resample our data to introduce this effect for study.

To study the impact of selection bias, we conduct additional experiments where we sample the original data to vary the correlation between age and Spanish, while holding constant the relationship of Spanish and callbacks. To increase the prevalence of Spanish among younger applicants, we drop at random older applicants with Spanish and younger applicants without Spanish, while keeping the callback rate of Spanish applicants constant. We consider a range of values for the conditional probabilities $P(\text{Spanish} = 1|A = y)$ and $P(\text{Spanish} = 1|A = o)$ to investigate how disparity in Spanish by age influences system behavior.

7 Results

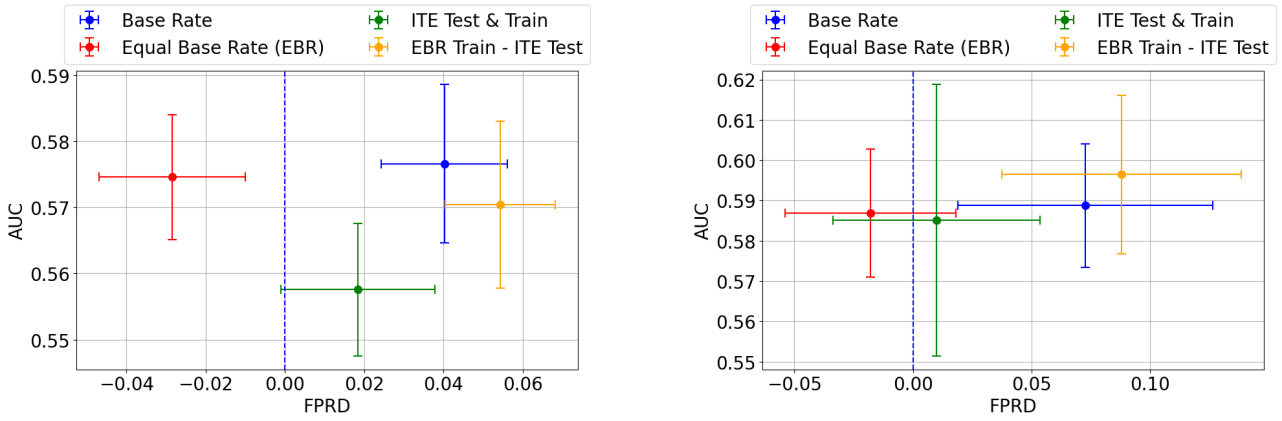
RQ1: Effect of label bias in test data. Figure 2 displays the main AUC versus FPRD results, with mean and standard deviation computed from 5-fold cross-validation. We first focus on the effects of debiasing the test data with ITE. For random forest (Figure 2a), the Base Rate has FPRD ≈ 0.038, exhibiting similar discrimination against older applicants as observed in the original dataset. Applying the Equal Base Rate intervention, and evaluating on the unmodified test set, at first appears to have removed the discrimination. Indeed, EBR seems to have over-adjusted, now showing a preference for older applicants (FPRD ≈ -0.031). However, we observe a dramatic difference when we repair the label bias in the test set. EBR Train - ITE Test shows that the discrimination against older applicants remains, even after equalizing base rates, and indeed the sign of FPRD changes when label bias is removed (FPRD ≈ 0.042 vs. $-.031$ for EBR). This discrepancy of 0.073 in FPRD suggests that label bias can have dramatic impacts on fairness.

For MLP (Figure 2b), the relative comparisons between the methods are largely similar. Noticeable differences are a somewhat higher AUC overall compared to RF (e.g., Base Rate $0.576 \rightarrow 0.59$), as well as much larger values for FPRD. For example, EBR Train - ITE Test increases FPRD from 0.042 using RF to 0.095 using MLP. This suggests that models with more degrees of freedom may be even more susceptible to label bias. Thus, in answering RQ1, we see that simply employing EBR can give the illusion of fairness, but in reality still exhibits an absolute $\approx 10\%$ age disparity.

RQ2: Effect of label bias in training data. Continuing the discussion of Figure 2, we next compare the result of ITE Test & Train, which performs the ITE label debiasing on both the training and test sets. This results in the lowest amount of discrimination (FPRD ≈ 0.017), though this does coincide with a roughly 1% point decrease in AUC. When compared with EBR Train - ITE Test, we see that ITE Test & Train reduces FPRD from $0.042 \rightarrow 0.017$, a 60% reduction in disparity. Even though EBR ensures that the two age groups receive equal callback rates in the training data, the label bias remains. Thus, under-qualified younger applicants receive callbacks at higher rates than under-qualified older applicants. And, conversely, qualified older applicants are more likely to not receive a callback than qualified younger applicants. By learning these patterns, the resulting classifier replicates this discrimination in the test set.

RQ3: Effect of label bias magnitude. To investigate the impact of the amount of human discrimination in the original audit data, we create a modified version of the data with more discrimination than the original. Specifically, we sample the original data by removing at random older applicants who received a callback until the callback difference between age groups increases to 10 percentage points, roughly double the original difference. Figure 3 shows the results for all methods. We can see that the overall patterns remain. Critically, we observe that the discrepancies in evaluation measures exhibit a commensurate increase in severity due to the doubling of discrimination. For example, for random forest (Figure 3a), EBR has FPRD $\approx -.038$, compared to an FPRD ≈ 0.133 for EBR Train-ITE Test. This is a discrepancy of 0.171, more than double the discrepancy of 0.073 in the original results in Figure 2a. Again, without adjusting for label bias, the sign of the discrimination is estimated incorrectly. These suggest that label bias can be more detrimental precisely in domains with large amounts of discrimination. This aligns with other existing studies on the effect of selection and label bias on synthetic data (Favier et al. 2023).

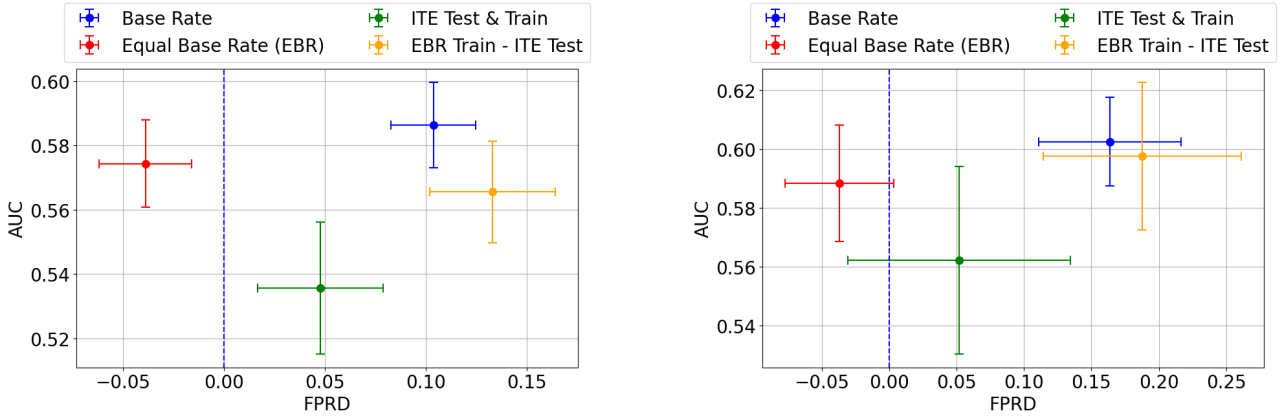
Encouragingly, ITE Test & Train appears to maintain only



(a) Random Forest (RF)

(b) Neural Network (MLP)

Figure 2: Accuracy (AUC) vs. fairness (FPRD) by method. FPRD > 0 indicates discrimination against older applicants.



(a) Random Forest (RF)

(b) Neural Network (MLP)

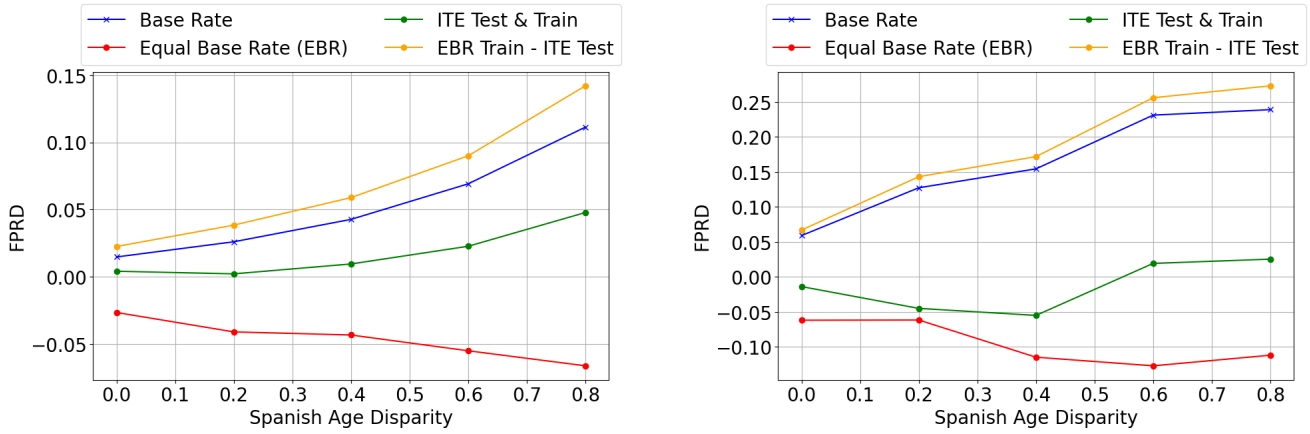
Figure 3: Accuracy (AUC) vs. fairness (FPRD) by method. Human discrimination in the audit data is doubled from that of Figure 2. FPRD > 0 indicates discrimination against older applicants.

modest discrimination even in this more extreme setting (FPRD $\approx$ 0.049 vs. 0.017 in the original), which is also observed in the neural network results (Figure 3b). While discrimination does grow, the lower starting point suggests the ITE approach can be more robust to higher levels of discrimination in the audit data. Unfortunately, this does appear to come at a decrease in AUC of roughly 2.1% points.

RQ4: Effect of selection bias. Finally, we analyze how results vary when we reintroduce selection bias into the dataset. Figure 4 plots FPRD as we vary Spanish disparity by age group. The x-axis $P(\text{Spanish} = 1|A = y) - P(\text{Spanish} = 1|A = o)$ shows much more likely younger applicants are to speak Spanish than older applicants. For example, when $x = 0.8$, $P(\text{Spanish} = 1|A = y) = 0.9$ and $P(\text{Spanish} = 1|A = o) = 0.1$. In the sample, 90% of younger vs. 10% of older applicants speak Spanish. With callback rates otherwise constant, Spanish being desirable leads to greater discrimination (higher FPRD) against older applicants as their Spanish-speaking share declines.

As in Figure 2, EBR differs markedly depending on whether the test set has been de-biased. This is most pronounced when Spanish Age Disparity is 0.8, where for EBR FPRD $\approx -.08$ and for EBR - ITE Test FPRD ≈ 0.145 for random forest (Figure 4a). The discrepancy is even larger for the neural network, which shows a FPRD discrepancy of 0.38 ($-0.11 \rightarrow 0.27$, Figure 4b). Here, the difference is not only large but also reverses direction from against younger applicants to against older applicants. E.g., for the MLP, two identically trained models can show to either have a 0.11 discrimination towards older applicants, or a 0.27 against younger ones, depending on whether test data is repaired.

Although the ITE Test & Train approach consistently achieves the lowest disparity on debiased test sets, it still exhibits a bias in favor of younger applicants under the most extreme conditions (e.g., at a bias level of 0.8). This further reinforces the value of audit study data, showing the limitations of relying solely on statistical corrections. By randomizing both covariates and protected attributes, audit studies



(a) Random Forest

(b) Neural Network

Figure 4: Fairness comparison as selection bias due to Spanish varies by age group. Spanish Age Disparity is $P(\text{Spanish} = 1|A = y) - P(\text{Spanish} = 1|A = o)$, i.e., larger x values mean larger Spanish in younger applicants.

offer a cleaner and more controlled source of training data that can mitigate the influence of sampling bias.

8 Discussion and Limitations

We demonstrate that achieving group parity in observed outcomes, e.g., callback rates, is an insufficient pre-processing intervention to ensure fairness. Interventions that equalize base rates but ignore label bias may create the *illusion of fairness*, leaving meaningful disparities unaddressed. This has consequences for both researchers and practitioners in domains like hiring, where there are interactions between features and the hiring process (Schumann et al. 2020).

In complex domains like hiring, feature interactions mean that resampling to equalize callback rates can obscure discrimination, especially when the original labels stem from biased decisions. A system that appears fair in terms of aggregate outcomes may still produce systematically different errors across groups. This underscores the need for fairness metrics and interventions that go beyond surface-level parity and consider the causal mechanisms of disparities. We show the importance of richer evaluation data – audit studies or experiments – that precisely identify when and how discrimination occurs. Without it, fairness assessments risk relying on biased labels and flawed assumptions.

Limitations. While our study provides a novel approach to evaluating fairness using audit study data and individual treatment effect estimation, several limitations remain. Audit studies better approximate discrimination, but like others they still lack access to the underlying ground truth of applicant quality. This limits our ability to definitively assess whether corrected labels fully reflect fair outcomes. Future work should investigate rigorous simulation studies to better understand how robust these approaches are to different distributions of label bias. Our analysis considers a limited set of fairness interventions and measures. While we demonstrate the shortcomings of base rate equalization and propose an ITE-based alternative, a larger comparison with other debiasing methods – e.g., adversarial learning, reweighting schemes, or post hoc calibration (Pessach and Shmueli 2022) – would provide a broader understanding of label bias. Finally, the ITE approach may introduce new biases depending on which instances it modifies. Hence expanding our framework to other techniques is a key step.

9 Conclusion and Future Work

We introduced a novel approach using human audit study data to better measure and mitigate algorithmic fairness in the presence of label bias. Our method leverages Individual Treatment Effect (ITE) estimates to assess whether individuals receive fair predictions and repairs the data accordingly. Our empirical results indicate that this approach leads both to fairer predictions can reduce the "*illusion of fairness*" of traditional approaches that do not account for label bias. These results point toward the need for future studies that can efficiently incorporate audit studies into AI-augmented decision making processes. Future studies should investigate rigorous simulation studies to test robustly how these approaches are to different distributions of label bias, including other protected attributes (e.g., race or gender). We evaluate our methods on a single dataset focused on age discrimination in hiring. This high-impact domain warrants future work across more datasets and settings to test generalizability. Our overall results underscore the importance of collecting data within the fair machine learning ecosystem, and not relying on simple data repair methods.

Acknowledgments

All authors are supported by the Tulane University Center of Excellence for Community-Engaged Artificial Intelligence; Sariola was also supported by the Harold L. and Heather E. Jurist Center of Excellence for Artificial Intelligence; Sariola, Culotta and Mattei also by CNS-SCC-2427237. Culotta was supported in part by the Louisiana Board of Regents Endowed Chairs for Eminent Scholars program.

References

Bandy, J. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Bogen, M.; and Rieke, A. 2018. Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias. Technical report, Upturn.

Burn, I.; Button, P.; Figinski, T. F.; and McLaughlin, J. S. 2020. Why Retirement, Social Security, and Age Discrimination Policies Need to Consider the Intersectional Experiences of Older Women. *Public Policy & Aging Report*, 30(3): 101–106.

Byun, Y.; Sam, D.; Oberst, M.; Lipton, Z.; and Wilder, B. 2024. Auditing Fairness under Unobserved Confounding. In Dasgupta, S.; Mandt, S.; and Li, Y., eds., *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, 4339–4347. PMLR.

Carey, A. N.; and Wu, X. 2022. The causal fairness field guide: Perspectives from social and formal sciences. *Frontiers in big Data*, 5: 892837.

Collins, J. C.; Chong, W. W.; De Almeida Neto, A. C.; Moles, R. J.; and Schneider, C. R. 2021. The simulated patient method: Design and application in health services research. *Research in Social and Administrative Pharmacy*, 17(12): 2108–2115. Publisher: Elsevier BV.

Corbett-Davies, S.; Gaebler, J. D.; Nilforoshan, H.; Shroff, R.; and Goel, S. 2023. The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24(312): 1–117.

Derous, E.; and Ryan, A. 2018. When your resume is (not) turning you down: Modelling ethnic bias in resume screening. *Human Resource Management Journal*, 29.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Evans, S. C.; Roberts, M. C.; Keeley, J. W.; Blossom, J. B.; Amaro, C. M.; Garcia, A. M.; Stough, C. O.; Canter, K. S.; Robles, R.; and Reed, G. M. 2015. Vignette methodologies for studying clinicians' decision-making: Validity, utility, and application in ICD-11 field studies. *International Journal of Clinical and Health Psychology*, 15(2): 160–170. Publisher: Elsevier BV.

Fan, J.; Upadhye, S.; and Worster, A. 2006. Understanding receiver operating characteristic (ROC) curves. *Canadian Journal of Emergency Medicine*, 8(1): 19–20.

Farber, H. S.; Herbst, C. M.; and Silverman, D. 2019. Whom Do Employers Want? The Role of Recent Employment and Unemployment Status and Age. *Journal of Labor Economics*, 37(2): 323–349.

Favier, M.; Calders, T.; Pinxteren, S.; and Meyer, J. 2023. How to be fair? A study of label and selection bias. *Machine Learning*, 112(12): 5081–5104.

Fawkes, J.; Fishman, N.; Andrews, M.; and Lipton, Z. 2024. The fragility of fairness: Causal sensitivity analysis for fair machine learning. *Advances in Neural Information Processing Systems*, 37: 137105–137134.

Fish, B.; Kun, J.; and Lelkes, Á. D. 2016. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM international conference on data mining*, 144–152. SIAM.

Foster, J. C.; Taylor, J. M.; and Ruberg, S. J. 2011. Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30(24): 2867–2880.

Gaddis, S. M. 2018. An introduction to audit studies in the social sciences. In *Audit studies: Behind the scenes with theory, method, and nuance*, 3–44. Springer.

Hardy, J. H.; Tey, K. S.; Cyrus-Lai, W.; Martell, R. F.; Olstad, A.; and Uhlmann, E. L. 2022. Bias in Context: Small Biases in Hiring Evaluations Have Big Consequences. *Journal of Management*, 48(3): 657–692.

Hutchinson, B.; and Mitchell, M. 2019. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, 49–58.

Jiang, H.; and Nachum, O. 2020. Identifying and correcting label bias in machine learning. In *International conference on artificial intelligence and statistics*, 702–712. PMLR.

Kilbertus, N.; Ball, P. J.; Kusner, M. J.; Weller, A.; and Silva, R. 2020. The Sensitivity of Counterfactual Fairness to Unmeasured Confounding. In Adams, R. P.; and Gogate, V., eds., *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, 616–626. PMLR.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Lahey, J. N. 2008. Age, Women, and Hiring: An Experimental Study. *Journal of Human Resources*, 43(1): 30–56.

Langer, M.; König, C.; Sanchez, D.; and Samadi, S. 2020. Highly automated interviews: applicant reactions and the organizational context. *Journal of Managerial Psychology*.

Langer, M.; König, C. J.; and Papathanasiou, M. 2019. Highly automated job interviews: Acceptance under the influence of stakes. *International Journal of Selection and Assessment*, 27(3): 217–234.

Li, N.; Goel, N.; and Ash, E. 2022a. Data-Centric Factors in Algorithmic Fairness. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, 396–410. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392471.

Li, N.; Goel, N.; and Ash, E. 2022b. Data-centric factors in algorithmic fairness. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 396–410.

Lippens, L.; Vermeiren, S.; and Baert, S. 2023. The state of hiring discrimination: A meta-analysis of (almost) all recent correspondence experiments. *European Economic Review*, 151: 104315.

Machado, A. F.; Charpentier, A.; and Gallic, E. 2025. Sequential conditional transport on probabilistic graphs for interpretable counterfactual fairness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18): 19358–19366.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Mitchell, S.; Potash, E.; Barocas, S.; D'Amour, A.; and Lum, K. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual review of statistics and its application*, 8(1): 141–163.

Neumark, D.; Burn, I.; and Button, P. 2019. Is It Harder for Older Workers to Find Jobs? New and Improved Evidence from a Field Experiment. *Journal of Political Economy*, vol. 127, no. 2].

Newman, D. T.; Fast, N. J.; and Harmon, D. J. 2020. When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes*, 160: 149–167.

Pearl, J. 2010. On a class of bias-amplifying covariates that endanger effect estimates. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI)*, 417–424.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Pessach, D.; and Shmueli, E. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3): 1–44.

Plečko, D.; Bareinboim, E.; et al. 2024. Causal fairness analysis: a causal toolkit for fair machine learning. *Foundations and Trends in Machine Learning*, 17(3): 304–589.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5): 688.

Schumann, C.; Foster, J.; Mattei, N.; and Dickerson, J. 2020. We need fairness and explainability in algorithmic hiring. In *International conference on autonomous agents and multiagent systems (AAMAS)*.

Steiner, P. M.; Atzmüller, C.; and Su, D. 2016. Designing Valid and Reliable Vignette Experiments for Survey Research: A Case Study on the Fair Gender Income Gap. *Journal of Methods and Measurement in the Social Sciences*, 7(2): 52–94.

Stredwick, J. 2005. Chapter 4 - Recruitment. In *Introduction to Human Resource Management (Second Edition)*, 116–160. Oxford: Butterworth-Heinemann, second edition edition. ISBN 978-0-7506-6534-6.

Vecchione, B.; Levy, K.; and Barocas, S. 2021. Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21. ACM.

Verma, S.; Ernst, M. D.; and Just, R. 2021. Removing biased data to improve fairness and accuracy. *CoRR*, abs/2102.03054.

Wick, M.; Panda, S.; and Tristan, J.-B. 2019. Unlocking Fairness: a Trade-off Revisited. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, 1171–1180.