More Data Can Lead Us Astray: Active Data Acquisition in the Presence of Label Bias

Yunyi Li, Maria De-Arteaga, Maytal Saar-Tsechansky

University of Texas at Austin
yunyi.li, dearteaga, maytal.saar-tsechansky @mccombs.utexas.edu

Abstract

An increased awareness concerning risks of algorithmic bias has driven a surge of efforts around bias mitigation strategies. A vast majority of the proposed approaches fall under one of two categories: (1) imposing algorithmic fairness constraints on predictive models, and (2) collecting additional training samples. Most recently and at the intersection of these two categories, methods that propose active learning under fairness constraints have been developed. However, proposed bias mitigation strategies typically overlook the bias presented in the observed labels. In this work, we study fairness considerations of active data collection strategies in the presence of label bias. We first present an overview of different types of label bias in the context of supervised learning systems. We then empirically show that, when overlooking label bias, collecting more data can aggravate bias, and imposing fairness constraints that rely on the observed labels in the data collection process may not address the problem. Our results illustrate the unintended consequences of deploying a model that attempts to mitigate a single type of bias while neglecting others, emphasizing the importance of explicitly differentiating between the types of bias that fairness-aware algorithms aim to address, and highlighting the risks of neglecting label bias during data collection.

Introduction

There is sufficient understanding that machine learning (ML) algorithms can easily replicate and even exacerbate societal biases. Abundant empirical evidence of biased ML systems has been found in a variety of domains, including criminal justice (Angwin et al. 2016), healthcare (Obermeyer et al. 2019), human resources (Pessach and Shmueli 2020), and content moderation (Sap et al. 2019; Davidson et al. 2017). In many of these domains, algorithms are making or supporting high-stakes, life-changing decisions. Awareness of the risks of biases in ML systems has led to an exploration of methods to mitigate those biases. A majority of the works that propose approaches to mitigate algorithmic bias measure bias at a group level (Mitchell et al. 2018; Verma and Rubin 2018) and have been focused on one of two categories. One line of work aims to use available data to train algorithms that yield better group fairness measures. This is typically achieved by imposing different types of fairness constraints (Bellamy et al. 2018; d'Alessandro, O'Neil, and LaGatta 2017), including preprocessing data transformations (Zemel et al. 2013; Louizos et al. 2015; Lum and Johndrow 2016; Adler et al. 2018; Turchetta, Berkenkamp, and Krause 2016; Del Barrio et al. 2018); in-processing optimization constraints (Woodworth et al. 2017; Zafar et al. 2017; Agarwal et al. 2018; Russell et al. 2017); and post-process group-specific classification thresholds (Feldman 2015; Hardt, Price, and Srebro 2016). The second approach aims to mitigate bias via additional data collection, which may be guided by a number of criteria, including the desire to cost-effectively produce a distribution that is expected to yield better generalization performance and smaller biases. For example, to prevent harm caused by ML systems, there have been numerous calls to obtain training datasets that are more inclusive, diverse, and representative of the populations of interest (Chen, Johansson, and Sontag 2018; Fazelpour and De-Arteaga 2022; Gebru et al. 2021; Mitchell et al. 2019; Veale and Binns 2017; Holstein et al. 2019), including calls by policy makers (European Commission 2021).

Because data collection–and especially label acquisition–is very costly, active learning is often used to assist the data collection process, so as to cost-effectively acquire labels that are particularly beneficial for learning (Saar-Tsechansky and Provost 2004; Saar-Tsechansky and Geva 2019). In contrast with traditional supervised learning, (pool-based) active learning is a framework in which the learner's goal is to proactively select a subset of examples to be labeled from a pool of unlabeled instances. Typically, an active learner is initially trained on a small set of labeled examples and adaptively determines the batch of unlabeled instances that would be most advantageous to be labeled next. Most active learning strategies choose instances to be labeled based on some notion of uncertainty, such as entropy of predictions (Shannon 1948). However, an instance's utility score (how beneficial it would be to include it in the training data) can be designed to depend on a number of criteria, and different heuristics can be used to achieve different performance needs. As recent work has noted, these heuristics can be modified to reflect fairness criteria. For example, if the goal is to learn a model that can mitigate the performance disparity across groups, then the active learning utility function can be designed to incorporate this objective (Abernethy et al. 2020). Thus, active learning's suitability for efficient and purpose-driven data collection has spurred interest at the intersection of algorithmic fairness and efficient data collection, and algorithms that propose active learning under fairness constraints have been developed. For instance, Abernethy et al. (2020) proposed adaptive sampling for min-max fairness by considering group membership when estimating the probability of sampling an instance to be labeled. Anahideh, Asudeh, and Thirumuruganathan (2022) proposed a method called "fair active learning" (FAL), which includes fairness improvements in the utility function in addition to overall model performance (eg. overall model accuracy).

Approaches for mitigating bias by imposing fairness constraints and existing approaches for fairness-oriented active data collection share one important commonality: they all assume (albeit often implicitly) that the observed label can be considered to be an unbiased "gold standard". Thus, the proposed bias mitigation strategies typically overlook bias present in the observed labels. This can be a problem because historical prejudice, bias, and unequal access to opportunities do not only affect covariates, but may also affect the labels used for training across domains. For example, Obermeyer et al. (2019) found that one widely used algorithm in the U.S. healthcare system meant to identify patients that will benefit most from enrollment in "high-risk care management" programs exhibited bias against Black patients due to label bias. The algorithm relied on healthcare costs as a proxy for health needs, but unequal access to care has historically led to lower healthcare spending for Black patients as compared to white patients, resulting in the algorithm underestimating their needs. Racial bias in data labels has also been shown to be present in multiple widely used Twitter corpora annotated for offensive language, where tweets inferred to be African American English are more likely to be annotated as offensive language (Sap et al. 2019). Research has also shown that models trained on these corpora have the potential to reinforce the racial bias and further marginalize voices of minorities (Sap et al. 2019).

In this work, we provide a structured overview of different types of label bias and empirically study the fairness implications of active data collection algorithms in the presence of label bias. Our work is grounded on a disconnect between the characteristics of common active learning settings, and the underlying assumptions of proposed bias mitigation strategies. Overlooking label bias in active data collection may be particularly problematic given that some of the most common active learning applications rely on human-generated labels, such as crowdsourced annotations (Yan et al. 2011), which are highly bias prone due to human cognitive bias (Haselton, Nettle, and Murray 2015). Furthermore, the settings used to motivate active learning with fairness constraints are often some in which label bias is a high risk. For instance, a recently proposed methodology termed "fair active learning" (FAL) (Anahideh, Asudeh, and Thirumuruganathan 2022) uses recidivism prediction as one of two motivating examples and evaluates the proposed approach in the COMPAS dataset, while recent work has highlighted the risk of label bias when using rearrest as a proxy for recidivism (Bao et al. 2021; Fogliato et al. 2021).

We study the impact of overlooking label bias during data collection empirically, using both simulations and real data. Our results show that if we overlook label bias while acquiring labels: 1) collecting more data can lead to exacerbated bias; 2) data-driven strategies to identify the "disadvantaged group" based on performance gaps can lead to misidentification; 3) relative comparisons of bias based on performance gaps across models can be misleading, which may misguide model selection.

In the next section, we briefly review extant research on algorithmic fairness (bias mitigation algorithms and calls for data collection), active learning, fairness-aware active learning, and label bias. Then we conceptualize label bias in the context of supervised learning systems and provide a structured overview of different types of label bias. We then describe our methodology for assessing the fairness performance of recently proposed fairness-aware active learning and the most commonly used active learning strategies in the presence of label bias. Thereafter, we present the results and show three *more data can lead us astray* patterns. We conclude by discussing the implications of our findings and opportunities for future research directions.

Related Work

In recent years, a large body of research has focused on the development of **bias mitigation algorithms**. Most of them fall into one of three streams: pre-processing, in-processing, or post-processing. Pre-processing methods try to remove information related to sensitive feature(s). The idea is that we first drop the sensitive feature(s), then learn a new feature space that removes the information correlated to the sensitive feature (Zemel et al. 2013; Louizos et al. 2015; Adler et al. 2018; Calmon et al. 2017). Alternatively, in-process methods try to modify the loss function in order to penalize the disparities in performance, for example, by adding a constraint or a regularization term to the existing loss function (Zafar et al. 2017; Gordaliza et al. 2019; Agarwal et al. 2018; Zafar et al. 2017; Woodworth et al. 2017; Calders, Kamiran, and Pechenizkiy 2009). Generally, the constraint is a quantitative operationalization of an underlying notion of fairness, and can be added to the objective function of any supervised ML model. The trade-off between the existing objective (eg. accuracy) and fairness metrics can be treated as a user parameter and thus can be adjusted based on different contexts and stakeholders' interests. Post-processing methods try to find appropriate thresholds using the original scoring function for each group (Hardt, Price, and Srebro 2016). As they are using the original scoring function, there is no need to retrain the model, thus is preferred if computation expense is an issue, but may come at a cost of other notions of fairness (Cheng et al. 2021). All of the aforementioned fairness-aware algorithms discuss the bias mitigation techniques with respect to observed labels, however, we consider label bias and differentiate observed labels from the labels of interest for algorithm training.

Despite these "fair" algorithm developments on fixed datasets, many researchers attribute the disparities to the unrepresentative of training samples (Buolamwini and Gebru 2018; Angwin et al. 2016), and advocate for **data collection** to avoid discrimination (Chen, Johansson, and Sontag 2018; Gebru et al. 2021; Mitchell et al. 2019; Veale and Binns 2017; Holstein et al. 2019). As we have mentioned in the introduction section, **active learning** approaches (Lewis and Gale 1994; Settles and Craven 2008; Settles 2009; Zhu et al. 2009; Huang, Jin, and Zhou 2010) are considered a cost-effective way to acquire additional training instances, since the algorithm can guide the selection of the most informative set of instances to be labeled and added to the training set. Driven by the need of collecting more data to mitigate bias cost-effectively, **fairness-aware active learning** methods have been proposed recently. Anahideh, Asudeh, and Thirumuruganathan (2022) proposed a query strategy that sample the next batch of instances considering both overall model information gain and fairness measure. At each iteration, the instance(s) with maximum Shannon entropy (Shannon 1948) and expected fairness improvement would be labeled and added to the training set. However, the obtained labels are still assumed to be the "gold standard". Similarly, other very recent works also incorporate the fairness notion in active learning strategy design (Abernethy et al. 2020; Sharaf and Daumé III 2020; Cai et al. 2022). All these new approaches inherit from classical active learning the assumption that the acquired label is a perfect match with the label of interest, but this assumption does not hold in a wide range of practical scenarios.

A separate body of work has devoted significant attention to the presence of **bias in human-generated labels**. With the flourishing of crowdsourcing services (Howe 2008; Yan et al. 2011), such as Amazon Mechanical Turk, the data labeling process is increasingly reliant on crowd work (Gray and Suri 2019). Crowd workers can perform as well as domain experts in certain tasks (Snow et al. 2008), especially when the composition of the workers' pool is carefully curated in a task-dependant manner (Allen et al. 2021). However, researchers have brought attention to the risk of annotators' cognitive bias (Eickhoff 2018; Draws et al. 2021), stereotyping encoded in annotators' assessments (Otterbacher 2015), and uneven representations of demographic characteristics among annotators (Barbosa and Chen 2019). A number of factors, including task and instructions clarity (Wu and Quinn 2017), task design (Kazai et al. 2011), incentives (Shah and Zhou 2015), and quality control mechanisms (Ipeirotis, Provost, and Wang 2010; McDonnell et al. 2016), have been demonstrated to affect the quality of the annotations (Draws et al. 2021). Furthermore, even when labels are collected from domain experts, this does not mean they are free of bias. For instance, in the context of healthcare, the quality of medical diagnoses and treatments in acute, cancer, and palliative pain care was compromised due to medical care providers' biases (Hoffman et al. 2016).

It is worth mentioning that there is a large body of literature that relaxes the assumption of perfect labels, focused on **learning from noisy labelers**. Researchers have proposed aggregating multiple noisy labelers' opinions either through majority voting (Zhang, Wu, and Sheng 2016), as well as learning probabilistic models to jointly estimate labelers' quality and gold standard labels (Snow et al. 2008; Smyth et al. 1994; Dawid and Skene 1979; Whitehill et al. 2009; Rodrigues, Pereira, and Ribeiro 2013; Welinder et al. 2010; Jin and Ghahramani 2002; Liu et al. 2012; Yan et al. 2010), or other heuristics (Huang et al. 2017; Gao and Saar-Tsechansky 2020). These proposed algorithms, nevertheless, assume constrained forms of noise such as random noise, which excludes shared societal biases, or dismiss fairness considerations in their model evaluation. Recent work has aimed to move beyond these assumptions by leveraging disagreement and recognizing annotators' unique perspectives (Davani, Díaz, and Prabhakaran 2022), but there is still a lack of active learning approaches that tackle this problem.

While label noise in ML systems has received extensive attention, there are relatively few works explicitly focused on **label bias**. Fogliato, Chouldechova, and G'Sell (2020) found that even small biases in observed labels is sufficient to lead to disparities in the recidivism risk assessment tool's disparate performance on different racial groups. The fact that "re-arrest"—the target variable used for training recidivism risk assessment tools—, is a different construct than "re-offend"—the outcome the risk assessment tool aims to predict—can cause fairness issues; "a model that appears to be fair with respect to rearrest could be an unfair predictor of re-offense" (Fogliato, Chouldechova, and G'Sell 2020). Label bias has been identified as a potential problem in other contexts such as child maltreatment hotline screenings (De-Arteaga, Dubrawski, and Chouldechova 2021), healthcare (Obermeyer et al. 2019), and offensive language detection (Sap et al. 2019). Obermeyer et al. (2019) and Sap et al. (2019) attribute the performance disparity of an ML system to a specific label bias under the context they examined, and De-Arteaga, Dubrawski, and Chouldechova (2021) propose a methodology that combines observed outcomes and human decisions to better approximate a construct of interest. Finally, the context-dependent and complex relationship between the observed labels used for training ML models, the construct of interest that an algorithm aims to predict, and the decision-making space has been discussed (Friedler, Scheidegger, and Venkatasubramanian 2016; Passi and Barocas 2019; Jacobs and Wallach 2021).

How Bias May Creep into Labels

In this section, we describe how bias may creep into instances' labels used to train ML models. Before we dissect different sources of label bias, let us define what we mean when we use the term *label bias*. Given an observed label $\tilde{Y}$ that is usually readily accessible in a set of training examples, a construct of interest Y^* that is an ideal label for training, and a variable denoting a binary group membership G, we refer to label bias as a systematic mismatch between the construct of interest and the observed label, such that the relationship underlying the mismatch differs across groups. $\tilde{Y}$ exhibits label bias with respect to Y^* if there is a group of relevance $g_0 \in G$ such that,

$$P(Y^* = \tilde{Y}|G = g_0) \neq P(Y^* = \tilde{Y}|G \neq g_0) \qquad (1)$$

For example, G may correspond to gender and g_0 to women. Label bias differs from label errors given its non-random nature and tight relationship with group fairness. It is also important to note that according to this definition, the presence of label bias cannot be determined on the basis of data alone; it is defined in reference to a construct of interest Y^* for training a supervised learning system, and thus depends on the predictive task that a supervised learning system is attempting to perform. Such bias may be a result of different phenomena, which we detail below.

Construct Gap occurs when there is a mismatch in theoretical definitions between the construct of interest and the construct of observed label (Jacobs and Wallach 2021). The reliance on a construct that does not match the construct of interest is typically motivated by the relative accessibility of one over the other. In some cases, this may occur because ML systems are often trained by repurposing previously collected datasets stored in organizational information systems. As a result, the goals during data collection may not match the goals during model development. For example, in the context of healthcare, financial incentives often result in detailed and meticulous data of insurance claims, which is then repurposed for multiple tasks in ML for healthcare.

The risks of relying on insurance claims and spending information as a proxy for healthcare conditions are illustrated in a study by Obermeyer et al. (2019) analyzing racial disparities in an algorithm used to prioritize patients that may benefit from care management programs meant for patients with complex health needs.

The algorithm used health costs as a proxy label for health needs. Due to historical inequities, in the US the costs incurred by Black patients are often considerably lower than those of white patients with similar healthcare needs. As a result, the algorithm prioritized healthier white patients over more ill Black patients.

Construct validity issues may also arise because high-level, complex objectives are often not directly quantifiable (Passi and Barocas 2019). As a result, it is necessary to choose proxies. Consider the task of predicting a student's likelihood to be successful in a college admission setting. It is impossible to fully capture the outcome of a 'successful student' using one single measurable feature due to the complex and potentially contentious definition of 'success'. In cases like this, a simplified construct such as 'GPA' or 'class ranking' is used, which may ignore various indicators of success, and inadvertently favor some population groups (Suresh and Guttag 2019).

Label Measurement Bias occurs when the construct we are interested in when training a supervised learning model is fully aligned with the construct we intended to measure, but measurement errors vary across groups (Jacobs and Wallach 2021). While this may also be thought of as a proxy problem, label measurement bias differs from construct gap in that there is no *conceptual* mismatch between the observed construct and the construct of interest. For example, measurement error from pulse oximetry is more prevalent for people with darker skin pigmentation. Compared to white patients, Black patients had roughly three times the rate of occult hypoxemia that was not identified by pulse oximetry (Sjoding et al. 2020). Thus, if an ML model is trained to predict oxygen saturation (Ghazal et al. 2019), the label we observe does match the construct we are interested in, but the measurement error is higher for Black patients as compared to white patients. In the context of ongoing efforts to use ML models to automatically adjust ventilation settings (Ghazal et al. 2019), this label measurement bias could result in misadjustment or delayed adjustment of ventilator settings for Black patients. Given that measurement often relies on various technologies as well as sociotechnical processes, disparities in measurement error are common across domains.

Human Labeling Bias arises in contexts when labels used for training ML models rely on human assessments, which is very common in crowdsourced data collections. In many tasks, ranging from radiology applications to misinformation detection, there is frequently no unique gold standard (Adamson and Welch 2019; Neumann, De-Arteaga, and Fazelpour 2022), and the construct of interest Y^* may not be directly observable and may be subject to disagreement (Aroyo et al. 2019). For example, when assessing candidates for job applications, committee members may disagree with each other on whether someone is qualified for the job. In this case, the labeling disagreements can be rooted in the differing subjectivities and value systems (Bless and Fiedler 2014). Crowd workers, whose life experience brings important perspectives to certain tasks may be dramatically underrepresented in generic crowdsourcing platforms such as Mturk, and the voice of the already unrepresentative group can be further marginalized (Davani, Díaz, and Prabhakaran 2022).

In addition to capturing a variety of perspectives, human-generated labels frequently capture societal biases and prejudices. Human labeling bias occurs when there is a mismatch between a construct of interest Y^* and the labels provided by human labelers $\tilde{Y}$, such that a group is systematically disfavoured. For example, (Otterbacher et al. 2018) found that people who are more sexist, as measured by the Ambivalent Sexism Inventory, are less likely to recognize and report gender biases in image search results, thereby reinforcing social stereotypes.

It is worth mentioning that these three types of label bias are not mutually exclusive. For example, construct gap and human labeling bias may often co-occur, as the proxy chosen may correspond to human assessments. For instance, when using income or promotions as a proxy for job skills or potential, there is both a construct gap issue and a risk of human labeling bias, as previous promotion and raise decisions are made by managers. However, conceptually distinguishing between the different types of label bias can help reason about the pathways through which these may capture societal injustices, and the ways in which the relationship between Y and Y^* may vary across groups.

Methodology

To empirically study the impact of overlooking label bias during a data collection process under (fairness-aware) active learning strategies, we use two datasets: the UCI Adult dataset (Dua and Graff 2017), and an offensive language dataset (Keswani, Lease, and Kenthapadi 2021). In this section, we provide a detailed description of the five (fairness-aware) active learning algorithms we empirically evaluated, the two datasets, and the experiment settings.

Algorithms Evaluated

We consider five active learning strategies: 1) Fair Active Learning; 2) Adaptive sampling; 3) Adaptive Sampling with uncertainty criteria; 4) Uncertainty Sampling; and 5) Random Sampling. The first three strategies are bias-mitigation active learning algorithms, the fourth strategy is a traditional and widely used active learning algorithm, and the last strategy is a naive sampling policy that is usually used as a baseline for the development of active learning algorithms. All of the five algorithms iteratively select a batch of unlabeled instances, query labels for them, and add them to the training data, but use different heuristics for the instance selection. Below we introduce the five different heuristics:

- *Fair active learning (FAL)* introduces group fairness constraints to traditional active learning objectives. It selects instances to be labeled based on the linear combination of two criteria: uncertainty-based Shannon entropy, and expected fairness improvement, measured using a group fairness metric (Anahideh, Asudeh, and Thirumuruganathan 2022). The penalty parameter controlling the trade-off between those two criteria is user input. Empirically, Anahideh, Asudeh, and Thirumuruganathan (2022) shows a considerable decrease in disparity while preserving accuracy when evaluating with respect to observed labels.
- *Adaptive sampling (Adaptive)* reduces disparate performance by either randomly selecting instances to be labeled from the disadvantaged group at a probability p, or randomly sampling instances from the full unlabeled pool at the probability $(1 - p)$ (Abernethy et al. 2020). The trade-off parameter p is a hyper-parameter, and the group considered the "disadvantaged group" is determined in a data-driven way at each iteration, defined as the group for which the algorithm has a lower performance at a given point.
- *Adaptive sampling with uncertainty criterion (Adapt. Uncert.)* is a natural extension to the original adaptive sampling method (Abernethy et al. 2020). We implement this variant by adding an uncertainty criterion (Settles 2009) to the sampling process. After deciding which group to sample from, instead of randomly sampling an instance, we use Shannon entropy (Shannon 1948) to select the instance that the algorithm is most uncertain about. This constitutes an active learning variant of adaptive sampling. Specifically, with probability p, we apply uncertainty sampling constrained to the "disadvantaged group", and with probability $(1-p)$ we apply uncertainty sampling to the full unlabeled pool of instances.
- *Uncertainty sampling (Uncertainty)* chooses the instances in each run to be labeled based on which instances the current model is most uncertain about (Settles 2009). A general method to measure uncertainty is Shannon entropy (Shannon 1948).
- *Random sampling (Random)* is commonly used as a baseline in the active learning literature. The method randomly selects an instance from the unlabeled pool.

Datasets

UCI Adult Dataset We use the Adult dataset (Dua and Graff 2017) from UCI Machine Learning Repository for income level prediction and we simulate a construct of interest for the dataset. The Adult dataset (Dua and Graff 2017), also known as the "census income" dataset, is extracted from the 1994 census data in the United States and is widely used for ML modeling and algorithmic fairness research. This dataset is a good example to represent construct gap bias, as the labels in the dataset—whether a given adult earns more than $50K per year—reflect historical inequalities that have resulted in lower wages for women and unpaid domestic labor. This in itself is not a problem with the data. But if income is assumed to be a proxy for other constructs of interest, such as contribution to the economy or deserving income based on skills, the construct gap can result in algorithmic bias. Labels in the dataset reveal dramatic economic disparities between men and women. Approximately one-third of men are reported to earn more than $50K per year, while only one-tenth of women are reported to have the same level of income. We use this dataset to conduct the first set of semi-synthetic experiments in which we simulate the relationship between $\tilde{Y}$ and Y^*, allowing us to have full control over this relationship in order to explore how label bias may affect active learning. For the simulation, we borrow the idea from Jiang and Nachum (2020)'s work and increased the female positive rate so that it matches the male positive rate. This scenario corresponds to a setting in which bias manifests itself as false negatives for women, with errors being independent of other observed variables. Specifically, we create a construct of interest, Y^*, by uniformly drawing a certain percentage of negative female instances and changing the label to be positive. This results in a label that satisfies statistical parity in the data, having the same proportion of each group labeled as positive.

Offensive Language Dataset Hate speech and offensive language identification are difficult because what is considered offensive depends on the social context. Derogatory terms toward African American communities have been reappropriated by these communities, and have gained different meanings and connotations in African American English (AAE), yet these terms remain offensive when used by outsiders (Sap et al. 2019). Thus, annotators may be more likely to classify an AAE tweet as offensive when it is not, a risk that may be exacerbated if annotators are not themselves familiar with AAE (Fazelpour and De-Arteaga 2022). We use a recently published offensive language dataset collected by Keswani, Lease, and Kenthapadi (2021), which contains 1471 Twitter posts, and is a subset of the 25k Twitter post curated by Davidson et al. (2017). Keswani, Lease, and Kenthapadi (2021) randomly assign the Twitter posts to 170 Amazon Mechanical Turk (Mtruk) labelers, so that each post is annotated by around 10 Mturk labelers regarding whether it contains offensive language. Thus, every post in Keswani,

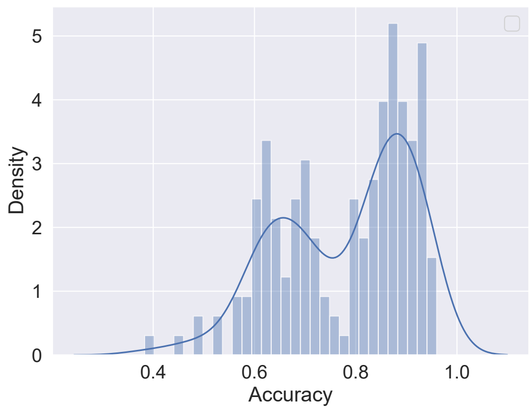

Figure 1: Offensive language annotation accuracy distribution of the 170 crowd labelers, evaluated on the "gold standard" labels. There is a considerable discrepancy on crowd workers' performance of accurately flagging the tweets that contain hate speech or offensive language.

Lease, and Kenthapadi (2021)'s dataset is associated with 1) a "gold standard" label that indicates whether it contains hate speech or offensive language based on Davidson et al. (2017)'s dataset; 2) a dialect feature that indicates whether the tweet's dialect is AAE. 3) around 10 newly acquired labels from different Amazon Mechanical Truk (Mturk) labelers.

Since the Mturk labelers were carefully recruited (Keswani, Lease, and Kenthapadi 2021), it is reasonable to assume that the labelers have the intention to provide accurate labels despite being potentially affected by biases and knowledge limitations. And because different people may have different knowledge and carry different biases based on their background and personal experience, labelers' performance also varies. Figure 1 shows the accuracy distribution of labelers assessed with respect to the original labels provided by Davidson et al. (2017). Our goal is to assess how observed biased labels may affect learning when we use (fair) active learning to acquire new training instances. To do this, we curate a set of labels that correspond to a "worst-case scenario" in which a single labeler is available per instance, and the available label corresponds to the labeler with the largest bias against AAE tweets. Specifically, we deem the labels provided by Davidson et al. (2017) as a gold standard, Y^*, and we use a subset of the labels collected by Keswani, Lease, and Kenthapadi (2021) to construct $\tilde{Y}$. For each instance, we let $\tilde{Y}$ be the label acquired from the labeler who exhibits the largest performance disparity when evaluated with respect to Y^*. It is worth mentioning that the labels in Davidson et al. (2017)'s dataset are majority votes of three or more crowdsourcing workers who were specifically instructed to label the tweet based on context, not the presence of particular words. Thus, we deem this as a gold standard because the instructions of the task are specifically designed to mitigate risks of bias stemming from ignoring context and over-relying on specific terms.

Experiment Setting

We train active learning algorithms using the two datasets with their biased labels, to mimic the situation of using active learning to acquire more data labels while ignoring label bias. We then evaluate model performance using both observed biased labels and the gold standards. Formally, let $\tilde{Y} \in \mathbb{R}^{n\times 1}$ and $Y^* \in \mathbb{R}^{n\times 1}$ be observed labels and gold standard, respectively. Let $X \in \mathbb{R}^{n\times d}$ be a matrix containing d attributes for n instances. We first split the observed dataset into training set $(X, \tilde{Y})_{\text{train}}$ and testing set $(X, \tilde{Y})_{\text{test}}$, and we use the training pool to perform active learning based on observed labels, $(X, \tilde{Y})_{\text{train}}$. Given a trained model f, we then obtain predictions $\hat{Y} = f(X_{\text{test}})$ on testing instances. Then, we evaluate $\hat{Y}_{\text{test}}$ with respect to both $\tilde{Y}_{\text{test}}$ and Y^*_{test}. We perform the experiments on 10 random train and test partitions of the dataset (70-30 split), and consider the mean and confidence interval over the 10 random splits. For the UCI Adult dataset, the maximum labeling budget was set at 200 (and we assume acquiring one label costs 1), after which the performance leveled out. Starting with 10 labeled instances (5 female and 5 male), we select one point to label at each active learning iteration in consecutive order until the budget is depleted. For the offensive language dataset, we start with 10 initial instances (5 AAE and 5 non-AAE) and acquire 10 instances' labels at each iteration. We specify the number of iterations to be 60 after which performance leveled off. We use the 100-dimensional vector representation of a tweet obtained from GloVe (Pennington, Socher, and Manning 2014) pre-trained word embedding as input and predict whether a tweet contains offensive language.

For active learning strategies with fairness considerations, such as adaptive sampling, adaptive uncertainty sampling, and FAL algorithm, we need to choose a fairness measure that we are interested in optimizing. We minimize the True Positive Rate (TPR) gap between female and male groups for income level prediction on the Adult dataset. as equalized opportunity (Hardt, Price, and Srebro 2016) is a sensible fairness goal for similar prediction tasks in human resource domains. We minimize the FPR gap between African American English (AAE) speakers and non-AAE speakers for offensive language detection tasks, as higher FPR for AAE tweets is the major inequality found in many offensive language detection algorithms (Sap et al. 2019).

Results and Analysis

In this section, we analyze the potential harms that can result from overlooking label bias in active label acquisition. To do this, we study the evolution of active learners (base models), comparing the *observed* evaluation metrics and the evaluation metrics with respect to the gold standard. We first show findings when performing simulations using the Adult dataset, followed by the findings when using the real-world offensive language dataset.

We found three outstanding patterns in both scenarios. If we overlook label bias while acquiring labels: 1) collecting more data can lead to exacerbated bias; 2) data-driven strategies to identify the "disadvantaged group" based on

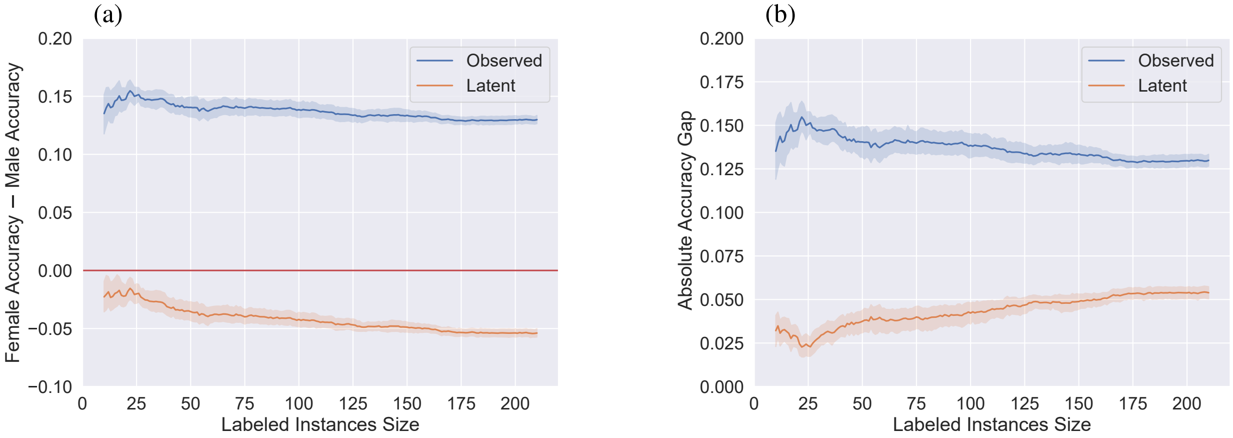

Figure 2: Gender accuracy gap (left plot) and absolute gender accuracy gap (right plot) versus number of newly acquired training instances. Shaded areas indicate 95% confidence interval. There are two insights: 1) misidentified disadvantaged group: evaluation on observed labels indicating male group to be the disadvantaged group, while in reality, female should be the disadvantaged group. 2) Bias may be interpreted as decreasing while the true bias is increasing.

performance gaps can lead to misidentification; 3) relative comparisons of bias across models can be misleading, which may misguide model selection.

More Data Can Exacerbate Bias

As we mentioned earlier in the introduction and related work, calls to mitigate bias by collecting more data have gained attention in recent years. In particular, it is expected that fairness-aware data collection can help address this problem (Chen, Johansson, and Sontag 2018). However, such strategies typically ignore potential label bias. We study the disparity between the bias mitigation performance evaluated on observed labels and the actual performance when evaluated on the gold standard labels. To accomplish so, we examine the accuracy (Acc) gaps and true positive rate (TPR) gaps, both in terms of the directed difference and the absolute gap. The directed accuracy gap which we used as one fairness metric is defined as the difference in accuracy between group g_1 and group g_0:

$$\begin{aligned}\text{AccGap} = &P(\hat{Y} = Y|G = g_0)\\ &-P(\hat{Y} = Y|G = g_1)\end{aligned} \tag{2}$$

Where $\hat{Y}$ and Y are random variables representing predicted and testing labels, and G is a random variable representing binary groups. $Y = \tilde{Y}$ when evaluated on observed labels and $Y = Y^*$ when evaluated on the construct of interest or the gold standard. As fairness metrics in proposed fairness-aware active learning algorithms often treat disparities as symmetric, we also visualize the accuracy gap as the absolute accuracy difference between g_1 and g_0:

$$\begin{aligned}\text{Abs. AccGap} = &|P(\hat{Y} = Y|G = g_0)\\ &-P(\hat{Y} = Y|G = g_1)|\end{aligned} \tag{3}$$

Similarly, we examine both the directed TPR gap and the absolute TPR gap between g_1 and g_0:

$$\begin{aligned}\text{TPRGap} = &P(\hat{Y} = Y|G = g_1, Y = 1)\\ &-P(\hat{Y} = Y|G = g0, Y = 1)\end{aligned} \tag{4}$$

$$\begin{aligned}\text{Abs. TPRGap} = &|P(\hat{Y} = Y|G = g_1, Y = 1)\\ &-P(\hat{Y} = Y|G = g_0, Y = 1)|\end{aligned} \tag{5}$$

In what follows, we present the results when applying the FAL algorithm for income prediction and applying uncertainty sampling for offensive language detection, which shows some of the starkest patterns of *more data can exacerbate bias*. The results of the other four active learning algorithms assessed in the paper show a similar pattern and can be found in the Appendix.

Income Prediction We begin by analyzing experiment results on the Adult dataset. Figure 2 (a) shows the dynamics of the female group's advantage in terms of accuracy as compared to the male group (AccGap) during the data collection process under the FAL algorithm, tested both on observed labels and on the (simulated) construct of interest. The x-axis indicates the number of newly acquired training instances using the FAL algorithm. The blue line represents AccGap tested on observed labels and the orange line represents AccGap tested on the construct of interest. The shaded area depicts a 95% confidence interval of the AccGap. Note that an accuracy gap of zero, visualized by the red horizontal line, marks the position where male accuracy and female accuracy are perfectly equal to each other. According to figure 2 (a), evaluation of observed labels shows that female accuracy is greater than male accuracy and data collection moderately mitigates this inequality. Meanwhile, evaluation of the construct of interest shows a very different pattern, indicating that female accuracy is lower than male accuracy, and more data leads to an enlarged accuracy gap and exacerbated the bias.

Figure 2 (b) illustrates the absolute accuracy gap. It is clear that the bias evolution trend is opposite when evaluated

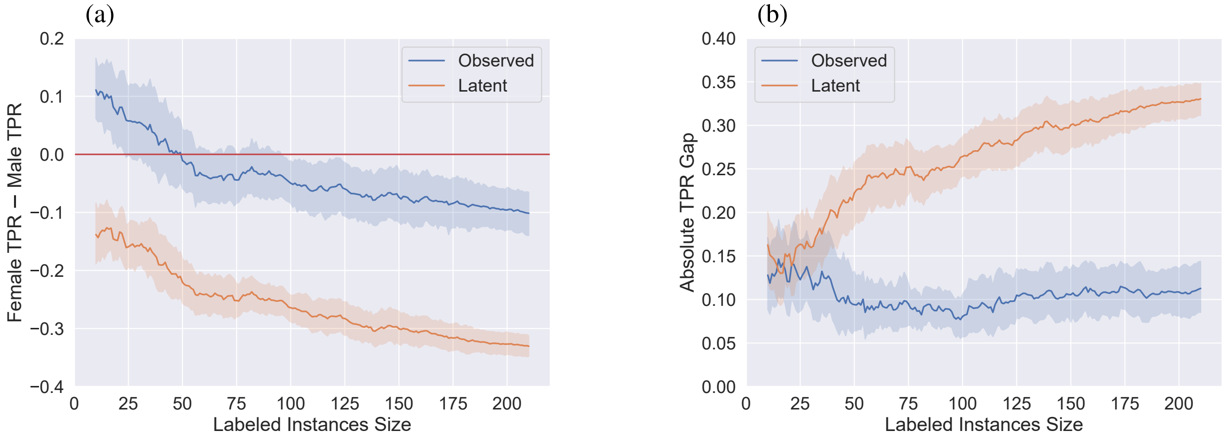

Figure 3: Gender TRP gap (left plot) and absolute gender TPR gap (right plot) versus the number of newly acquired training instances. Shaded areas indicate a 95% confidence interval. Evaluation of observed labels shows data collection mitigates the gender TPR gap if we carefully choose the stopping point, say if we stop around 80 newly labeled instances. However, in reality, collecting more data further exacerbated the existing gender TPR gap.

on the observed labels versus on the construct of interest. According to observed labels, the bias is reduced during the data collection process under the FAL framework, which aligns with the goal of collecting data to improve fairness. However, evaluation of the gold standard label indicates that the bias is actually aggravated as more data is collected.

Figure 3 illustrates the TPRGap of the FAL algorithm during the data collection process, tested on observed labels (blue line) and construct of interest (orange line). According to figure 3 (a), if label bias is overlooked, the evaluation on observed labels indicates female TPR started as greater than male TPR and the data collection mitigates the inequality and may over-correct the bias when newly collected instances size greater than 80. In such case, one may claim that the data collection is useful to mitigate bias. However, in reality, female TPR is initially less than male TPR, and collecting more data exacerbated this inequality. Similarly, the visualization of the absolute TPRGap in figure 3 (b) shows the anonymous and symmetric bias as TPRGap is slightly decreased and then remains stable based on observed labels, whereas the bias continues to aggravate in reality.

Offensive Language Detection We find a similar *more data can exacerbate bias* pattern in offensive language detection experiments based on a real-world dataset. As we described in the offensive language dataset section, racial bias is a central issue in offensive language detection tasks, considering the sensitive feature as indicating whether the dialect is African American English (AAE) or not. Let non-AAE be g_0 and AAE be g_1. Figure 4 illustrates the dynamics of the non-AAE group's advantage in terms of accuracy as compared to the AAE group during the data collection process under uncertainty sampling. It is clear that while the observed accuracy difference is approaching 0 as we acquire more data, the latent accuracy difference is getting larger. If we ignore the bias in the labels and depend just on the observed testing results to assess bias reduction performance, we might conclude that data collection aids bias mitigation, when in fact it exacerbates the problem.

Misidentifying Disadvantaged Group

Many works on algorithmic fairness propose the idea of automatically identifying the disadvantaged group by evaluating a measure of interest for different groups, then mitigating the bias based on the evaluation. For example, adaptive sampling iteratively evaluates and identifies the disadvantaged group, then samples new instances from that group. Similarly, FAL relies on selecting the instances that are most helpful for mitigating the gap of a measure of interest, which implicitly identifies the disadvantaged group and tries to improve the model's performance on the group that exhibits the worse performance (according to some fairness measure) at a given point.

Data-driven approaches to identify who the disadvantaged group is ignoring contextual factors and historical inequities. This can be particularly problematic when such inequities result in label bias, in which case there is a risk of misidentifying who the disadvantaged group is, resulting in counterproductive mitigation strategies. Figures 2, 3, 4 all demonstrate the risk of misidentifying the disadvantaged group.

Income Prediction In the context of income prediction, we know that women have historically had less access to opportunities and high salaries, and have had lower income levels than men. The Adult dataset verifies the pattern as the ratio of males earning more than $50K per year is three times higher than that of females in the dataset. In our experiments, it can be seen that if the observed label has a biased relationship with a latent construct of interest, such as the one we simulate, a reliance on the biased observed labels could lead females to be misidentified as the advantaged group, aggravating existing disparities.

Offensive Language Detection When considering offensive language detection, researchers have found that "AAE

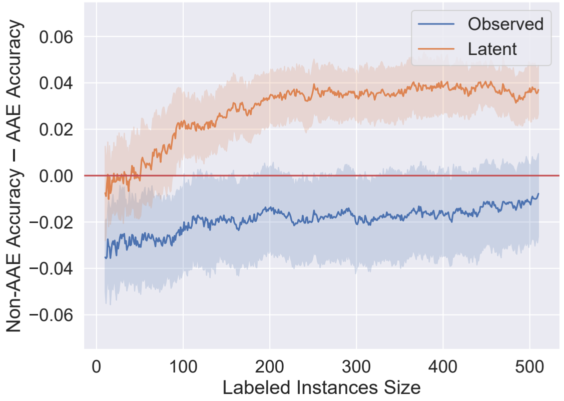

Figure 4: Accuracy difference between non-AAE tweets and AAE tweets versus the number of newly acquired training instances evaluated on observed labels (blue line) and the gold standards (orange line). The shaded area indicates a 95% confidence interval. Evaluation of observed labels indicates non-AAE accuracy is less than AAE accuracy and data collection can mitigate the inequality. However, evaluation of the gold standard shows that non-AAE accuracy is greater than AAE accuracy and data collection exacerbated this inequality.

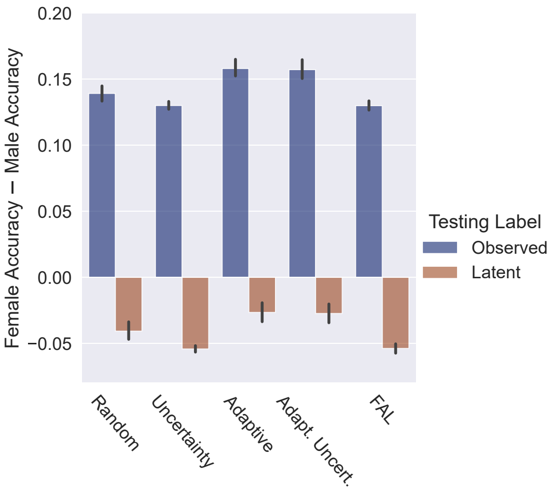

Figure 5: Gender accuracy gap for all five active learning algorithms evaluated on observed labels (blue bars) and construct of interest (orange bars). The vertical line in the middle of each bar illustrates a 95% confidence interval. Model selection can be misguided: evaluation on observed labels indicates uncertainty sampling to be the fairest algorithm whereas it is the least fair algorithm in reality.

tweets and tweets by self-identified African Americans are up to two times more likely to be labeled as offensive compared to others" (Sap et al. 2019). However, relying on evaluation on biased human labeling may result in treating AAE as the advantaged group, further marginalizing African Americans' voices.

Most proposed bias mitigation algorithms are anonymous and symmetric, a property that is often lauded as a favorable trait, since the application of such algorithms requires minimal contextual knowledge. However, with the prevalence of label bias, those algorithms run a risk of misidentifying the disadvantaged group, and thus mitigation can become exacerbation. Many fairness-aware active learning algorithms, such as adaptive sampling, rely on identifying the disadvantaged group at each iteration of the data collection process in order to collect data from the disadvantaged group; thus, misidentifying the disadvantaged group can violate the original intention of the active data acquisition method to reduce disparities. Less obviously but importantly, algorithms such as FAL actively select instances with two goals: improving overall accuracy and reducing non-directional fairness violations. In this case, implicitly misidentifying the disadvantaged group would misguide the algorithm to select instances that can improve the advantaged group's performance.

Mislead Model Selection

In addition to misguiding people's beliefs on bias mitigation performance of (fairness-aware) active learning assisted data collection, overlooking label bias can mislead model selection as well. Suppose that multiple algorithms are considered, and fairness metrics are used to select the data acquisition strategy that provides the best bias mitigation effect from a list of algorithms, given comparable accuracy.

Income Prediction Figure 5 illustrates the accuracy difference between female and male groups in income prediction for all five active learning frameworks we described in the 'Methodology' section, evaluated on observed labels (blue bars) and the gold standards (orange bars). By comparing the male accuracy minus female accuracy evaluated by biased observed labels (blue bars), we would select uncertainty sampling as the accuracy difference is the smallest among all strategies. However, the selected model, uncertainty sampling, could be the most biased model according to the evaluation of the construct of interest. Therefore, if we overlook label bias, model selection can be misleading and compound existing bias.

Offensive Language Detection The same pattern can be found in offensive language detection experiments as well. Figure 6 illustrates accuracy difference between AAE and non-AAE tweets for all five active learning algorithms. Based on the evaluation metrics uncertainty sampling could be selected as the best algorithm for bias mitigation, while it may actually be the most biased algorithm with respect to the gold standard label.

Discussion

As supervised learning algorithms are increasingly used for guiding decision-making in various high-stakes domains, and recognizing its reliance on potentially biased sources of labels, there is a growing need to understand the potential harm of label bias on (algorithm-assisted) data collection. In this paper, we presented an overview of different types of

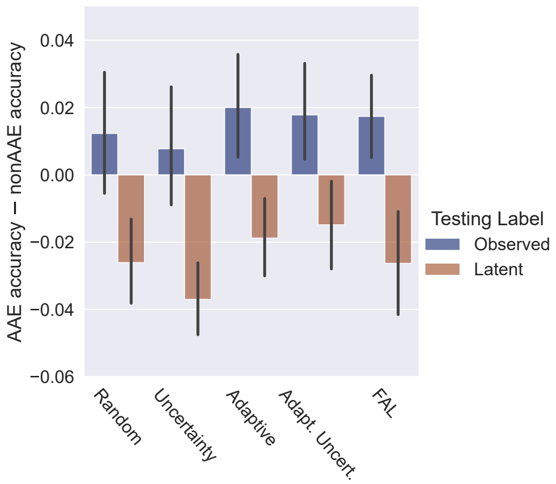

Figure 6: Accuracy difference between AAE and non-AAE for all five active learning algorithms evaluated on observed labels (blue bars) and the gold standards (orange bars). The vertical line in the middle of each bar illustrates a 95% confidence interval. Model selection can be misguided: evaluation on observed labels indicates uncertainty sampling to be the fairest algorithm whereas it is the least fair algorithm in reality.

label bias in the context of supervised learning systems and conducted empirical studies that uncover the potential effect of label bias on (fair) active learning algorithms. We evaluated the bias mitigation performance of the most commonly used active learning algorithms and the recently proposed fairness-aware active learning strategies using a combination of simulations and real-world data. Our study has demonstrated that if we overlook label bias while acquiring labels: 1) collecting more data can lead to exacerbated bias; 2) Using data-driven strategies to identify the "disadvantaged group" based on performance gaps can lead to misidentification; and 3) relative comparisons of bias across models can be misleading, which may misguide model selection.

Most of the existing work on algorithmic fairness overlooks label bias, and research emphasizing the risk of label bias has been primarily conceptual, with only a few studies providing empirical evidence of risks. To the best of our knowledge, our work is the first to study how label bias may mislead active data collection, and how the introduction of fairness constraints that overlook label bias may fail to address the problem. In addition to showing how data collection can be led astray by label bias, our work has implications for the discussion regarding the identification of "disadvantaged groups". A growing body of research has proposed methods that automatically identify disadvantaged groups, e.g. (Abernethy et al. 2020), and the "symmetric" idea of bias is a common property of algorithmic fairness algorithms that consider any disparities in error rates to be indicative of bias. Our work provides empirical evidence showing that this can be problematic and may inadvertently exacerbate harm to already marginalized groups, since label bias may lead to a misidentification of the groups that need protection or are harmed by disparities.

When we say more data may lead to exacerbated bias, this does not mean that data collection efforts are inherently unproductive. Rather, our findings underscore the need for developing fairness-aware active learning algorithms that consider label bias. More broadly, this work highlights the importance of developing methodologies for bias detection and mitigation in the presence of label bias. Given the current state of the art, ML practitioners may mitigate the risks we uncover by checking the underlying assumptions of proposed algorithms, differentiating the types of bias that fairness-aware algorithms aim to address, and scrutinizing the data generating process to better understand the potential issues in the data.

Limitations and Future Work

Label bias is often highly context-specific. The patterns we found in this study may only represent a portion of the possible consequences of ignoring label bias. We encourage future works that explore other possible harms in different application domains and under different assumptions of the relationship between observed labels and constructs of interest. In particular, in our experiments using the Adult dataset, which relies on simulations, we have made simplifying assumptions regarding the relationship between observed labels and gold standard labels. This by no means aims to provide a faithful estimation of a specific construct, such as economic contribution or skills. We use this as an example to empirically study the potential risks of label bias in a simplified setting.

In the context of offensive language data, we rely on real data for both gold standard labels and observed labels, which introduces the challenge that both labels may be flawed. The gold standard is more reliable than what we assume to be the observed labels due to (1) the instructions of the task that we consider the gold standard, which specifically aim to mitigate risks of bias stemming from ignoring context and over-relying on specific terms; (2) our construction of the "observed pool" as containing the assessments of those who exhibit the largest disparities. This is useful to empirically study how inadvertently relying on biased annotations could mislead the data collection process. However, this provides an assessment with respect to a point that may itself contain some bias, underestimating the magnitude of the problem.

Additionally, the harms we study have been focused on a definition of label bias grounded on group fairness. Considering label bias in relation to other definitions of fairness, such as individual fairness, may shed light on patterns and risks that are not visible through a group fairness lens. Similarly, our empirical evaluation is limited to statistical disparities and does not consider dynamic, long-term harms (Liu et al. 2018). In some cases, data collection may involve the allocation of goods and burdens, and thus studying the dynamic effects may visibilize additional risks that are beyond the scope of this study.

Acknowledgements

This research was supported by a Google AI Award for Inclusion Research, the Machine Learning Laboratory[1], and Good Systems[2], a UT Austin Grand Challenge to develop responsible AI technologies.

References

Abernethy, J.; Awasthi, P.; Kleindessner, M.; Morgenstern, J.; and Zhang, J. 2020. Adaptive Sampling to Reduce Disparate Performance. *arXiv preprint arXiv:2006.06879*.

Adamson, A. S.; and Welch, H. G. 2019. Machine learning and the cancer-diagnosis problem-no gold standard. *The New England journal of medicine*, 381(24): 2285–2287.

Adler, P.; Falk, C.; Friedler, S. A.; Nix, T.; Rybeck, G.; Scheidegger, C.; Smith, B.; and Venkatasubramanian, S. 2018. Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54(1): 95–122.

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*, 60–69. PMLR.

Allen, J.; Arechar, A. A.; Pennycook, G.; and Rand, D. G. 2021. Scaling up fact-checking using the wisdom of crowds. *Science advances*, 7(36): eabf4393.

Anahideh, H.; Asudeh, A.; and Thirumuruganathan, S. 2022. Fair active learning. *Expert Systems with Applications*, 199: 116981.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *'https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing'*.

Aroyo, L.; Dixon, L.; Thain, N.; Redfield, O.; and Rosen, R. 2019. Crowdsourcing subjective tasks: the case study of understanding toxicity in online discussions. In *Companion proceedings of the 2019 world wide web conference*, 1100–1105.

Bao, M.; Zhou, A.; Zottola, S.; Brubach, B.; Desmarais, S.; Horowitz, A.; Lum, K.; and Venkatasubramanian, S. 2021. It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. *arXiv preprint arXiv:2106.05498*.

Barbosa, N. M.; and Chen, M. 2019. Rehumanized crowdsourcing: A labeling framework addressing bias and ethics in machine learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Bellamy, R. K.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; et al. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

Bless, H.; and Fiedler, K. 2014. *Social cognition: How individuals construct social reality*. Psychology Press.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.

Cai, W.; Encarnacion, R.; Chern, B.; Corbett-Davies, S.; Bogen, M.; Bergman, S.; and Goel, S. 2022. Adaptive Sampling Strategies to Construct Equitable Training Datasets. *arXiv preprint arXiv:2202.01327*.

Calders, T.; Kamiran, F.; and Pechenizkiy, M. 2009. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, 13–18. IEEE.

Calmon, F. P.; Wei, D.; Vinzamuri, B.; Ramamurthy, K. N.; and Varshney, K. R. 2017. Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 3995–4004.

Chen, I.; Johansson, F. D.; and Sontag, D. 2018. Why is my classifier discriminatory? *Advances in Neural Information Processing Systems*, 31.

Cheng, M.; De-Arteaga, M.; Mackey, L.; and Kalai, A. T. 2021. Social Norm Bias: Residual Harms of Fairness-Aware Algorithms. *arXiv preprint arXiv:2108.11056*.

d'Alessandro, B.; O'Neil, C.; and LaGatta, T. 2017. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data*, 5(2): 120–134.

Davani, A. M.; Díaz, M.; and Prabhakaran, V. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10: 92–110.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Dawid, A. P.; and Skene, A. M. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1): 20–28.

De-Arteaga, M.; Dubrawski, A.; and Chouldechova, A. 2021. Leveraging expert consistency to improve algorithmic decision support. *arXiv preprint arXiv:2101.09648*.

Del Barrio, E.; Gamboa, F.; Gordaliza, P.; and Loubes, J.-M. 2018. Obtaining fairness using optimal transport theory. *arXiv preprint arXiv:1806.03195*.

Draws, T.; Rieger, A.; Inel, O.; Gadiraju, U.; and Tintarev, N. 2021. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, 48–59.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository. *University of California, Irvine, School of Information and Computer Sciences*.

Eickhoff, C. 2018. Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining*, 162–170.

[1] https://ml.utexas.edu/

[2] http://goodsystems.utexas.edu/

European Commission. 2021. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts.

Fazelpour, S.; and De-Arteaga, M. 2022. Diversity in sociotechnical machine learning systems. *Big Data & Society*, 9(1): 20539517221082027.

Feldman, M. 2015. *Computational fairness: Preventing machine-learned discrimination*. Ph.D. thesis.

Fogliato, R.; Chouldechova, A.; and G'Sell, M. 2020. Fairness evaluation in presence of biased noisy labels. In *International Conference on Artificial Intelligence and Statistics*, 2325–2336. PMLR.

Fogliato, R.; Xiang, A.; Lipton, Z.; Nagin, D.; and Chouldechova, A. 2021. On the Validity of Arrest as a Proxy for Offense: Race and the Likelihood of Arrest for Violent Crimes. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 100–111.

Friedler, S. A.; Scheidegger, C.; and Venkatasubramanian, S. 2016. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*.

Gao, R.; and Saar-Tsechansky, M. 2020. Cost-accuracy aware adaptive labeling for active learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2569–2576.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Ghazal, S.; Sauthier, M.; Brossier, D.; Bouachir, W.; Jouvet, P. A.; and Noumeir, R. 2019. Using machine learning models to predict oxygen saturation following ventilator support adjustment in critically ill children: A single center pilot study. *PloS one*, 14(2): e0198921.

Gordaliza, P.; Del Barrio, E.; Fabrice, G.; and Loubes, J.-M. 2019. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, 2357–2365. PMLR.

Gray, M. L.; and Suri, S. 2019. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29: 3315–3323.

Haselton, M. G.; Nettle, D.; and Murray, D. R. 2015. The evolution of cognitive bias. *The handbook of evolutionary psychology*, 1–20.

Hoffman, K. M.; Trawalter, S.; Axt, J. R.; and Oliver, M. N. 2016. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences*, 113(16): 4296–4301.

Holstein, K.; Wortman Vaughan, J.; Daumé III, H.; Dudik, M.; and Wallach, H. 2019. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–16.

Howe, J. 2008. *Crowdsourcing: How the power of the crowd is driving the future of business*. Random House.

Huang, S.-J.; Chen, J.-L.; Mu, X.; and Zhou, Z.-H. 2017. Cost-Effective Active Learning from Diverse Labelers. In *IJCAI*, 1879–1885.

Huang, S.-J.; Jin, R.; and Zhou, Z.-H. 2010. Active learning by querying informative and representative examples. *Advances in neural information processing systems*, 23: 892–900.

Ipeirotis, P. G.; Provost, F.; and Wang, J. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, 64–67.

Jacobs, A. Z.; and Wallach, H. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 375–385.

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.

Jin, R.; and Ghahramani, Z. 2002. Learning with multiple labels. In *NIPS*, volume 2, 897–904. Citeseer.

Kazai, G.; Kamps, J.; Koolen, M.; and Milic-Frayling, N. 2011. Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 205–214.

Keswani, V.; Lease, M.; and Kenthapadi, K. 2021. Towards Unbiased and Accurate Deferral to Multiple Experts. *arXiv preprint arXiv:2102.13004*.

Lewis, D. D.; and Gale, W. A. 1994. A sequential algorithm for training text classifiers. In *SIGIR'94*, 3–12. Springer.

Liu, L. T.; Dean, S.; Rolf, E.; Simchowitz, M.; and Hardt, M. 2018. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, 3150–3158. PMLR.

Liu, Q.; ICS, U.; Peng, J.; and Ihler, A. 2012. Variational inference for crowdsourcing. *sign*, 10: j2Mi.

Louizos, C.; Swersky, K.; Li, Y.; Welling, M.; and Zemel, R. 2015. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*.

Lum, K.; and Johndrow, J. 2016. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077*.

McDonnell, T.; Lease, M.; Kutlu, M.; and Elsayed, T. 2016. Why is that relevant? collecting annotator rationales for relevance judgments. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 4, 139–148.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, 220–229.

Mitchell, S.; Potash, E.; Barocas, S.; D'Amour, A.; and Lum, K. 2018. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*.

Neumann, T.; De-Arteaga, M.; and Fazelpour, S. 2022. Justice in Misinformation Detection Systems: An Analysis of Algorithms, Stakeholders, and Potential Harms. *arXiv preprint arXiv:2204.13568*.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.

Otterbacher, J. 2015. Crowdsourcing stereotypes: Linguistic bias in metadata generated via gwap. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1955–1964.

Otterbacher, J.; Checco, A.; Demartini, G.; and Clough, P. 2018. Investigating user perception of gender bias in image search: the role of sexism. In *The 41st International ACM SIGIR conference on research & development in information retrieval*, 933–936.

Passi, S.; and Barocas, S. 2019. Problem formulation and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 39–48.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Pessach, D.; and Shmueli, E. 2020. Algorithmic fairness. *arXiv preprint arXiv:2001.09784*.

Rodrigues, F.; Pereira, F.; and Ribeiro, B. 2013. Learning from multiple annotators: distinguishing good from random labelers. *Pattern Recognition Letters*, 34(12): 1428–1436.

Russell, C.; Kusner, M.; Loftus, C.; and Silva, R. 2017. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in neural information processing systems*, volume 30. NIPS Proceedings.

Saar-Tsechansky, M.; and Geva, T. 2019. More for less: Adaptive labeling payments in online labor markets. *Data Mining and Knowledge Discovery*, https://doi.org/10.1007/s10618-019-00637-z.

Saar-Tsechansky, M.; and Provost, F. 2004. Active sampling for class probability estimation and ranking. *Machine learning*, 54(2): 153–178.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, A. N. 2019. The risk of racial bias in hate speech detection. In *ACL*.

Settles, B. 2009. Active learning literature survey. *University of Wisconsin-Madison Department of Computer Sciences*.

Settles, B.; and Craven, M. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 1070–1079.

Shah, N. B.; and Zhou, D. 2015. Double or nothing: Multiplicative incentive mechanisms for crowdsourcing. *Advances in neural information processing systems*, 28: 1–9.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423.

Sharaf, A.; and Daumé III, H. 2020. Promoting fairness in learned models by learning to active learn under parity constraints. In *Workshop on Real World Experiment Design and Active Learning. International Conference on Machine Learning*.

Sjoding, M. W.; Dickson, R. P.; Iwashyna, T. J.; Gay, S. E.; and Valley, T. S. 2020. Racial bias in pulse oximetry measurement. *New England Journal of Medicine*, 383(25): 2477–2478.

Smyth, P.; Fayyad, U.; Burl, M.; Perona, P.; and Baldi, P. 1994. Inferring ground truth from subjective labelling of venus images. In *Advances in neural information processing systems*, volume 7.

Snow, R.; O'connor, B.; Jurafsky, D.; and Ng, A. Y. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, 254–263.

Suresh, H.; and Guttag, J. V. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*.

Turchetta, M.; Berkenkamp, F.; and Krause, A. 2016. Safe exploration in finite markov decision processes with gaussian processes. *Advances in Neural Information Processing Systems*, 29: 4312–4320.

Veale, M.; and Binns, R. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2): 2053951717743530.

Verma, S.; and Rubin, J. 2018. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, 1–7. IEEE.

Welinder, P.; Branson, S.; Perona, P.; and Belongie, S. 2010. The multidimensional wisdom of crowds. *Advances in neural information processing systems*, 23: 2424–2432.

Whitehill, J.; Wu, T.-f.; Bergsma, J.; Movellan, J.; and Ruvolo, P. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, 22: 2035–2043.

Woodworth, B.; Gunasekar, S.; Ohannessian, M. I.; and Srebro, N. 2017. Learning non-discriminatory predictors. In *Conference on Learning Theory*, 1920–1953. PMLR.

Wu, M.-H.; and Quinn, A. 2017. Confusing the crowd: Task instruction quality on amazon mechanical turk. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 5, 206–215.

Yan, Y.; Rosales, R.; Fung, G.; and Dy, J. G. 2011. Active learning from crowds. In *ICML*.

Yan, Y.; Rosales, R.; Fung, G.; Schmidt, M.; Hermosillo, G.; Bogoni, L.; Moy, L.; and Dy, J. 2010. Modeling annotator expertise: Learning when everybody knows a bit of something. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 932–939. JMLR Workshop and Conference Proceedings.

Zafar, M. B.; Valera, I.; Rogriguez, M. G.; and Gummadi, K. P. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, 962–970. PMLR.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In *International conference on machine learning*, 325–333. PMLR.

Zhang, J.; Wu, X.; and Sheng, V. S. 2016. Learning from crowdsourced labeled data: a survey. *Artificial Intelligence Review*, 46(4): 543–576.

Zhu, J.; Wang, H.; Tsou, B. K.; and Ma, M. 2009. Active learning with sampling by uncertainty and density for data annotations. *IEEE Transactions on audio, speech, and language processing*, 18(6): 1323–1331.