When Fair Ranking Meets Uncertain Inference

Avijit Ghosh
Northeastern University
avijit@ccs.neu.edu

Ritam Dutt
Carnegie Mellon University
rdutt@andrew.cmu.edu

Christo Wilson
Northeastern University
cbw@ccs.neu.edu

ABSTRACT

Existing fair ranking systems, especially those designed to be demographically fair, assume that accurate demographic information about individuals is available to the ranking algorithm. In practice, however, this assumption may not hold — in real-world contexts like ranking job applicants or credit seekers, social and legal barriers may prevent algorithm operators from collecting peoples' demographic information. In these cases, algorithm operators may attempt to infer peoples' demographics and then supply these inferences as inputs to the ranking algorithm. This raises the issue that errors in inference may subvert the fairness objectives that the ranking algorithm is attempting to guarantee.

In this study, we investigate how uncertainty and errors in demographic inference impact the fairness offered by fair ranking algorithms. Using simulations and two case studies with real datasets, we show how demographic inferences drawn from real systems can lead to unfair rankings. Our results suggest that operators should approach the use of inferred demographic data with caution, and further highlight the social and legal tensions between data collection and the use of fair-by-design algorithms.

ACM Reference Format:
Avijit Ghosh, Ritam Dutt, and Christo Wilson. 2021. When Fair Ranking Meets Uncertain Inference. In *FAccT '21: Conference on Fairness, Accountability, and Transparency (FAccT '21), March 2021, Toronto, Ontario, Canada.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.XXXX/XXXXXX.XXXXXX

1 INTRODUCTION

Social biases in algorithms have been investigated and identified in a number of contexts [16, 18, 25, 30, 34, 35, 39, 43, 51, 55–57]. As a result, there is now a thriving research community that seeks to develop fair algorithms [6, 13, 26, 31, 68] and efforts by activists and regulators to see that these tools are adopted in practice [20].

However, there is an underlying assumption with the vast majority of existing fair algorithms that ground-truth information about "protected" classes is available to the algorithm. This data is crucial, as it is used to measure and control for social bias, thus enabling fair outcomes. In cases where people are the data subjects being input to classification or ranking algorithms, it is assumed that demographic information will be available to mitigate sexism, racism, ageism, and other social biases. Unfortunately, this assumption about the availability of ground-truth data is often violated in practice. For example, in real-world contexts like automatically assessing job applicants or credit seekers, social and legal barriers may prevent algorithm operators from collecting peoples' demographic information.

The unavailability of ground-truth data has led some system developers to adopt an alternative approach: infer protected class information, such as demographics, from data, and then supply the inferred data to the fair algorithm as input. One example of this is the Bayesian Improved Surname Geocoding (BISG) inference algorithm that is used by lenders and health insurers in the U.S. to infer people's race and ethnicity [1, 12]. This demographic data is used to ensure that lenders are making race-neutral lending decisions and that health insurers are not discriminating based on race. Given the high-stakes of these use cases, it is clear that accurate demographic information is critical, lest unchecked discrimination lead to serious harm for individuals.

The use of inferred data raises the issue that errors in inference may subvert the fairness objectives that a fair algorithm is attempting to optimize for. If the underlying information about protected class status contains a significant number of errors, the fair algorithm cannot be expected to control for underlying social biases, since those biases may no longer be evident in the data. To the best of our knowledge, this problem has not been explored systematically in the literature, despite the fact that consequential real-world systems like BISG have already adopted the practice.

In this study, we investigate how uncertainty and errors in demographic inference impact fairness guarantees in the context of ranking algorithms. We approach this question using two complementary techniques. *First*, we use simulations to explore the relationship between population demographics, fairness metrics, and errors in inference. Simulations allow us to precisely specify the breakdown of the population (i.e., the number of protected groups and their relative size) and the demographic inference error rates, so that we can measure how these variables impact four different measures of ranking fairness. However, while simulations enable experimental control. *Second*, to address the issue of ecological validity, we examine two case studies based on real-world datasets (COMPAS [1] and the Adult Income Dataset[2]). Each of these datasets includes ground-truth demographic data, which enables us to generate a baseline unfair ranking and an "optimal" fair ranking (with respect to a specific fair ranking algorithm and objective function). We then compare these lower and upper bounds against rankings generated by a fair ranking algorithm when using erroneous demographic inferences as input. To further increase the ecological validity of our case studies, we present results using

FAccT '21, March 2021, Toronto, Ontario, Canada

ACM ISBN XXX-XXX-XXX…$15.00
https://doi.org/10.XXXX/XXXXXX.XXXXXX

[1] https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis
[2] https://archive.ics.uci.edu/ml/datasets/adult

demographic inference error rates drawn from five real-world algorithms, including BISG.

Our results suggest that developers should approach the use of inferred demographic data with caution. Our simulations and case studies make clear that errors in inference can dramatically undermine the purported guarantees of fair ranking algorithms, causing them to produce rankings that are much closer to the unfair baseline than the optimal fair ranking. We even observe instances where erroneous inferences cause groups that were not particularly disadvantaged in the baseline unfair ranking to become disadvantaged after a "fair" re-ranking.

We conclude this paper by discussing the implications of our findings. While it may be tempting to approach these issues through the lens of technical fixes, e.g., improved demographic inference, or uncertainty-aware fair ranking algorithms, we argue that these approaches have serious limitations, as well as disturbing normative implications. We also draw attention to the fundamental tension between privacy and fair algorithms, in that the solution to erroneous inference is to use high-quality ground-truth demographic data. However, achieving this may require changes to the law as well as peoples' habits.

2 BACKGROUND

We begin by briefly discussing fairness issues with machine learning algorithms in general, and ranking algorithms in particular, as they are the focus of our study. We also touch on documented issues with demographic-based classification and inference.

Algorithmic Fairness. Hype around big data and "artificial intelligence" has sharpened concerns about the social impact of these systems. Barocas and Selbst [4] and Osoba and Welser IV [52] talk about the different ways in which these algorithms may perpetuate societal biases across a variety of contexts. To mitigate these problems, research is being pursued in two directions — firstly, the discovery of biases in algorithms via academic toolkits like LIME [54] and SHAP [42] or corporate toolkits like AIF360 [5] and Google's What-IF tool [64]. Secondly, there has been work on fairness-aware machine learning algorithms, like in classification [24, 29, 32, 46], regression [2, 6], causal inference [41, 49], word embeddings [9, 10], machine translation [21] and finally, ranking [15, 61, 69].

Fair Ranking. Several fair ranking algorithms have been proposed in the literature. Early approaches, like the framework by Celis et al. [15] and FA*IR by Zehlike et al. [69], are binary optimizers, meaning that they only attempt to make a ranked list fair between two groups — protected and unprotected. More recent methods use constrained learning and treat the fair ranking as an optimization problem [61]. There are also other approaches, like achieving fairness through pairwise comparisons [7] and applying fairness constraints in learning-to-rank methods [47, 70].

Fair Ranking Metrics. Mehrabi et al. [45] provide a comprehensive list of the different fairness definitions that appear in the fair machine learning literature. Drawing from both classical and statistical notions of fairness, there exist concepts like *equalized odds* and *equal opportunity* [26], *demographic parity*, *treatment parity* [19], etc.

These concepts have been adapted specifically to the domain of ranking – there are several metrics in the literature to measure the fairness of a ranking system's output with respect to different classes or groups of people or items that the system was sorting. Fairness as a concept is hard to crystallize into an objective value, and therefore it is up for debate what constitutes the "best" metric, especially given that what is "best" may change depending on the context. Metrics developed by Yang and Stoyanovich [65] measured the underlying population representation in the top-ranked items, while other metrics such as those by Singh and Joachims [61] and Sapiezynski et al. [59] conceptualized ranking fairness as an attention or exposure allocation problem to the different subgroups. Singh and Joachims [61] propose that attention allocation metrics correspond to the problem of *disparate impact*, i.e., since top-ranked results gain more visual attention than results at the bottom.

Another consideration is the cardinality of the protected categories themselves. Several metrics from earlier literature, such as those by Zehlike et al. [69] and Kuhlman et al. [38], are binary, i.e., they are only able to assess fairness between two classes or groups. Binary metrics cannot be used in situations where intersectional fairness is desired, e.g., fairness between White males and White females, as opposed to males and females without respect to other demographic traits. Newer metrics like those proposed by Geyik et al. [23] that compare entire population distributions over an unspecified number of subgroups, or attention-based metrics [8, 59, 61] that also deal with the population distributions are agnostic to group cardinality, and thus do not have this problem.

Problematically, a meta-analysis of these (and other) fair ranking metrics has shown that they often disagree about whether a given ranked list is "fair" [53]. This motivated us to choose several metrics for our study, as we describe in § 3.2.2.

Inferred Attributes. There are examples in the literature that highlight accuracy problems with demographic inference algorithms. Buolamwini and Gebru [11] showed how the accuracy of commercial facial analysis systems at predicting gender fell when presented with images of darker-skinned people. Kosinski and Wang [37] developed a (problematic and disputed) system for predicting sexual preferences from photos of faces, but the overall accuracy was fairly poor.

Complex interactions between noise in protected attribute data and algorithms trying to ensure fairness to all groups, protected an unprotected, is sparsely studied despite its potentially far-reaching consequences. There have been studies on the stability of classification algorithms with noisy data [58]. Friedler et al. [22] note that classifiers may not be stable in the face of variations in the training dataset. Celis et al. [14] present a framework to achieve fair classification under a significant amount of noise in inferred protected attributes. However, to the best of our knowledge, no work has looked at how noisy or imperfectly inferred protected attributes impact fair ranking.

3 ALGORITHMS AND METRICS

Before delving into the specific algorithms we used in our experiments, we first review the goal of this study and use it to motivate the requirements for selecting these algorithms.

In this study, we aim to investigate what happens to ranking fairness metrics when erroneous demographic inferences are taken as input by fair ranking algorithms. Rather than approaching this question theoretically, we do so empirically using simulations and case studies. Thus, we require several basic components to implement these empirical experiments, including: (1) one or more fair ranking algorithms to evaluate, (2) fair ranking metrics that encompass a spectrum of fairness definitions, and (3) error rates drawn from inference algorithms. To improve confidence in our results, we strive to adapt algorithms and datasets that are drawn from real-world deployments.

With these goals and guiding principles in mind, we now move on to selecting algorithms and metrics.

3.1 Fair Ranking Algorithm

The first major decision we needed to make to facilitate our study was choosing one or more fair ranking algorithms. Recall that our goal is to assess how well this algorithm or these algorithms are able to achieve their fairness objectives when given input data that includes ground-truth and inferred demographic information.

The fair ranking algorithm we chose for this work was an algorithm developed by Geyik et al. [23] at LinkedIn. This algorithm was applicable to our use case since we aimed to include >2 groups in our experiments. The paper presents four different re-ranking algorithms with varying stability but with one central goal: to achieve the desired distribution of population in the top-ranked results with respect to one or more protected attributes, while providing ways to tailor the target distribution to achieve various fairness criteria. At a high-level, the algorithm takes an unfairly arranged list and an integer K then generates a fairness-aware list of the top K candidates such that the fraction of candidates in each subgroup matches their fraction in the underlying population. While other algorithms from prior work have a similar goal, this algorithm was extensively tested and deployed in LinkedIn's Talent Search system. The authors of the paper claim that the deployment led to "tremendous improvement in the fairness metrics (nearly three-fold increase in the number of search queries with representative results) without affecting the business metrics, which paved the way for deployment to 100% of LinkedIn Recruiter users worldwide" [23]. Since our work focuses on the possible breakdown of fair ranking algorithms in real-world, deployed scenarios where protected attributes may not be completely available, this work was the best fit for our research purposes.

Of the four algorithms presented in the paper, we chose the Deterministic Constrained Sorting algorithm or *DetConstSort* as our benchmark fairness algorithm since it is theoretically proven to be feasible for protected attributes having a large number of possible attribute values. Other greedy fair ranking algorithms in the paper were proven to be infeasible if the possible number of attributes exceeded three. *DetConstSort* does not suffer from this weakness and hence was employed for this work.

DetConstSort creates a ranked list of candidates, such that for any particular rank k and for any group attributes g_j, the attribute occurs at least $\lfloor p_{g_j}.k \rfloor$ times in the ranked list. However, unlike other fair ranking algorithms that greedily pick the best candidate for a particular rank, the *DetConstSort* algorithm also strives to improve the sorting quality by re-ranking the candidates that come above it (so that candidates with better scores are placed higher in the list), as long as the resultant list satisfies the feasibility criteria. Thus, the algorithm can be conceptualized as solving a more general interval constrained sorting problem. Since the *DetConstSort* algorithm is constrained to be feasible it optimizes the Skew and NDKL fairness metrics, which we introduce in the next section.

3.2 Metrics for Fairness Evaluation

The second major decision we needed to make to accomplish our study was choosing a metric or metrics for evaluating the fairness of representation in ranked lists. Specifically, we focus on metrics that (1) attempt to assess *group fairness* as their baseline criteria, possibly balanced against secondary objectives, and (2) are capable of dealing with multiple subgroups and intersectional fairness (i.e., not just binary protected versus unprotected classes).

3.2.1 Representation-based Metrics. To get an overall sense of the group fairness is a given ranked list, we chose two representation-based metrics that are slightly modified forms of the metrics introduced by Geyik et al. [23]. These metrics do not incorporate attention, i.e., they assess the representation of people from different groups based solely on how many of those people appear in the list relative to the underlying population. The first metric is computed per group, while the latter is aggregated across groups.

Given a ranked list τ, the Skew for attribute value g_i at position k is defined as

$$\text{Skew}_{g_i}@k(\tau) = \frac{p_{\tau^k,g_i}}{p_{q,g_i}} \tag{1}$$

where p_{τ^k,g_i} represents the proportion of members belonging to group g_i within the top k items in the ranked list τ, and p_{q,g_i} represents the proportion of members belonging to group g_i in the overall population q. Ideally, $\text{Skew}_{g_i}@k$ should be close to one for each g_i and k, indicating that people from g_i are represented in τ proportionally relative to the underlying population. $Skew_{g_i}@k > 1$ denotes that the group g_i is overrepresented among the top K candidates, and vice versa when the $Skew_{g_i}@k < 1$

Given a ranked list τ, the Normalized Discounted Kullback–Leibler (KL) Divergence (NDKL) is defined as

$$\text{NDKL}(\tau) = \frac{1}{Z}\sum_{i=1}^{|\tau|}\frac{1}{log_2(i+1)}d_{KL}(D_{\tau^i}||D_r) \tag{2}$$

where $d_{\text{KL}}(D_1||D_2) = \sum_j D_1(j)log_2\frac{D_1(j)}{D_2(j)}$ is the KL divergence score of distribution D_1 with respect to distribution D_2 and $Z = \sum_{i=1}^{|\tau|}\frac{1}{log_2(i+1)}$. NDKL can be interpreted as a weighted average of the logarithm of the Skew scores for all the groups in a ranked list. NDKL values close to zero indicate that people from all groups are represented proportionally in a given ranked list, since the KL-Divergence of the population between the top K candidates and the underlying population will be zero. A large difference in the distributions of the different groups in the top K ranked candidates leads to a higher NDKL score.

3.2.2 Attention-based Metrics. Studies have repeatedly shown that people do not pay equal attention to all items in ranked lists[48, 50];

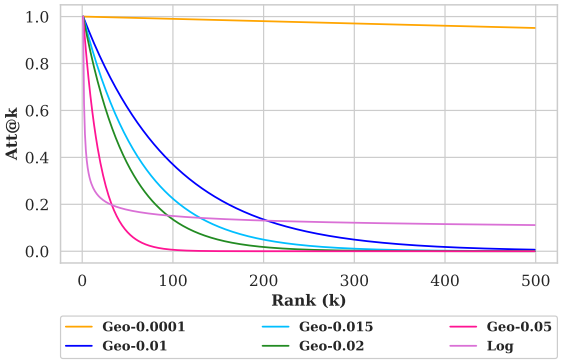


Figure 1: Attention versus rank for six attention functions.

rather, peoples' attention tends to decay as they progress down the list, eventually abandoning the task entirely even if more items are available. This observation suggests that using overall representation to assess fairness is misleading, since (1) people may not look at all available items and (2) they pay more attention and are thus more likely to act on higher ranking items. To take attention into account, we adopt six metrics that weigh tradeoffs between fairness and utility (i.e., relevance) in different ways. As in § 3.2.1, three metrics are computed per group and three are aggregated versions of the former across groups.

In this study, we adopted the geometric distribution to model decay in attention, similar to prior work by Sapiezynski et al. [59]. We compute attention at k as

$$\text{Att}_p@k(\tau) = 100 \times (1-p)^{k-1} \times (p) \tag{3}$$

where τ is the ranked list and p represents the proportion of attention provided to the first result. For our experiments we set $p = 0.015$ because at this value attention decays to zero at $k = 300$, which is the value of k we fix for our experiments. Although most prior work in the information retrieval literature uses logarithmic decay to model attention [61, 65], we did not adopt it because it models attention decay at an unrealistically slow rate and its shape flattens out at low ranks. Figure 1 shows how attention decays as a function of rank for a variety of values of p, as well as under logarithmic decay.

The i^{th} element in τ has an associated score, denoted s_i^τ, that corresponds to the utility or relevance of the item, and a group-attribute value, denoted by g_i^τ. The elements in the ranked list are arranged in decreasing order of score such that $s_i^\tau \geq s_j^\tau \; \forall \; i \leq j$. We define

$$\eta_{g_j,\tau} = \frac{1}{|g_j|} \sum_{i=1}^{|\tau|} \text{Att}_p@i \text{ where } g_i^\tau = g_j \tag{4}$$

where $\eta_{g_j,\tau}$ denotes the mean attention score of the g_j protected attribute for τ and

$$\text{ABR}_\tau = \frac{\min_{g_j}(\eta_{g_j,\tau})}{\max_{g_j}(\eta_{g_j,\tau})} \tag{5}$$

where ABR_τ or the **A**ttention **B**ias **R**atio for the ranking τ quantifies the disparity between the groups with the lowest and highest mean attention score ($\eta_{g_j,\tau}$). $\text{ABR}_\tau = 1$ is the ideal score, i.e., all groups receive equal attention.

In a similar fashion to attention score, we define $U_{g_j,\tau}$ to denote the mean utility score of the g_j protected attribute for τ.

$$U_{g_j,\tau} = \frac{1}{|g_j|} \sum_{i=1}^{|\tau|} s_i^\tau \text{ where } g_i^\tau = g_j \tag{6}$$

We define a metric θ_{τ,g_j} for a specified group g_j under a given ranking τ. It is defined as the ratio of the mean attention score to the mean utility score for the g_j under τ:

$$\theta_{\tau,g_j} = \frac{\eta_{g_j,\tau}}{U_{g_j,\tau}} \tag{7}$$

This construction is motivated by the *Disparate Treatment Constraint* in Singh and Joachims [61], with the idea being that in a fair ranking, the attention provided to an item should be proportional to its utility. We further define DTBR_τ or **D**isparate **T**reatment **B**ias **R**atio

$$\text{DTBR}_\tau = \frac{\min_{g_j}(\theta_{\tau,g_j})}{\max_{g_j}(\theta_{\tau,g_j})} \tag{8}$$

for a ranking τ, which quantifies the disparity between the groups with the lowest and highest θ_{τ,g_j}. As with ABR_τ, $\text{DTBR}_\tau = 1$ is the ideal score.

The probability of an action (e.g., a click or a view) being taken on an item in a ranked list is directly proportional to both the utility of that item and the attention provided to that item because of its position in the list. Similar to the *Click Through Rate* defined in Singh and Joachims [61], we define *action rate* α as:

$$\alpha_i = \text{Att}_p@i * s_i \tag{9}$$

Motivated by the *Disparate Impact Constraint* in Singh and Joachims [61], we define a metric γ_{τ,g_j} that is the mean action rate α_i for group g_i in τ.

$$\gamma_{g_j,\tau} = \frac{1}{|g_j|} \sum_{i=1}^{|\tau|} \alpha_i^\tau \text{ where } g_i^\tau = g_j \tag{10}$$

Finally, $DIBR_\tau$ or **D**isparate **I**mpact **B**ias **R**atio for a ranking τ quantifies the disparity between the groups with the lowest and highest γ_{τ,g_j}.

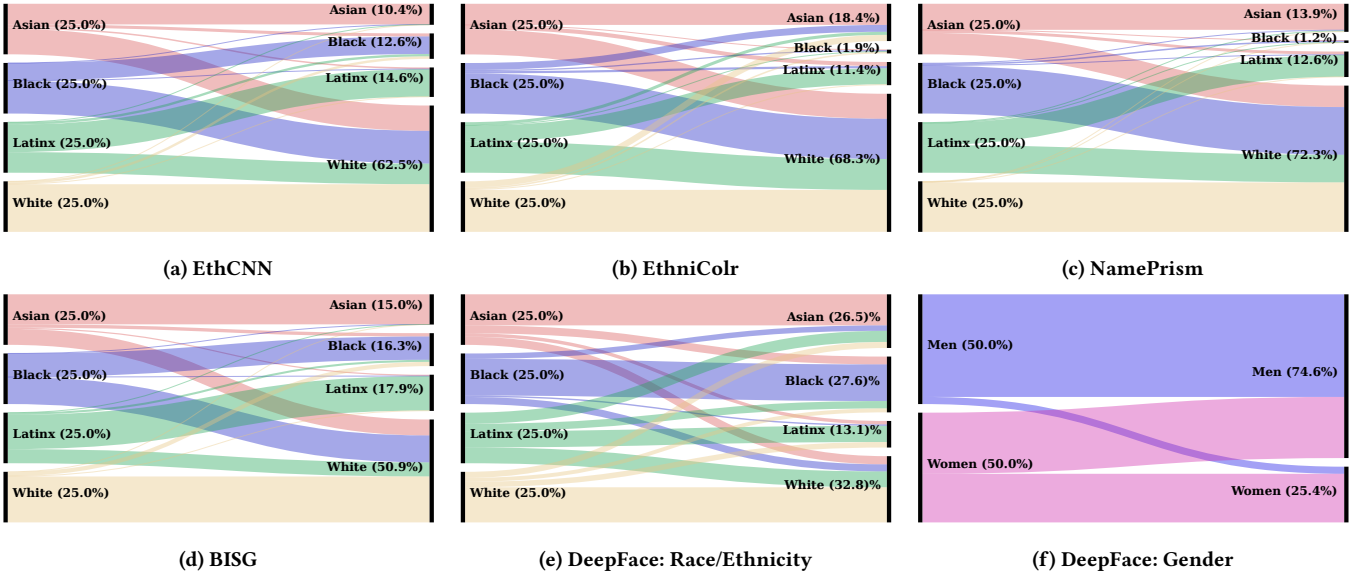
$$\text{DIBR}_\tau = \frac{\min_{g_j}(\gamma_{\tau,g_j})}{\max_{g_j}(\gamma_{\tau,g_j})} \tag{11}$$

As with our other metrics, $\text{DIBR}_\tau = 1$ is the ideal score.

3.3 Demographic Inference Algorithms

The third and final major decision we needed to make for this study was selecting demographic inference algorithms. Our intent is to compare the fair rankings generated by the *DetConstSort* algorithm when given ground-truth and inferred demographic information, using the metrics introduced in § 3.2, so as to quantify the impact (if any) of mis-classifications.

We chose five diverse inference algorithms that rely on different features and machine learning techniques. For each algorithm, we computed its confusion matrix when predicting peoples' ethnicity/race and gender (in one case) using ground-truth data with known demographics. We evaluated the four algorithms in



(a) EthCNN (b) EthniColr (c) NamePrism

(d) BISG (e) DeepFace: Race/Ethnicity (f) DeepFace: Gender

Figure 2: Sankey plots showing the distribution of ground-truth (left) and inferred (right) demographic traits for five algorithms. The algorithms tend to mis-classify minorities as Whites. DeepFace also mis-classifies Women to be Men a majority of cases.

§ 3.3.1 and § 3.3.2 using voter records from the state of North Carolina,[3] which are publicly available records that have the name, address, race, gender, and other personal information of each registered voter in the state. We evaluated the facial analysis algorithm in § 3.3.3 using the FairFace dataset [33].

Using these five algorithms we predicted race/ethnicity (Asian, Black, Latinx, and White) and gender (man and woman). We fully acknowledge that these categories are problematic, however, we adopted them because they are the categories supported by the inference algorithms from prior work. We unpack the many problems and limitations that derive from these categories in § 6.2.

Figure 2 shows the results of demographic inference using these five algorithms. We used these confusion matrices in our experiments to intentionally mis-classify data, so as to observe the effect on fair ranking performance.

3.3.1 Name-based Inference Algorithms. We chose two algorithms that attempt to predict peoples' race/ethnicity based on their name.

EthCNN. We employed a Convolutional Neural Network (CNN) architecture similar to Kim [36] to infer peoples' ethnicity from their names. We represented each individual's name as a sequence of characters c_i from c_1 to c_n, where n denoted the number of characters in the given name, with optional 'PAD' tokens to facilitate batching. We mapped each character into a 64 dimensional latent embedding, generated at random, and fine-tuned during the course of training. We applied a convolution filter of size k on the k concatenated character embeddings $c_{i:i+k}$, followed by non-linear activation function (ReLU). This yielded a scalar value s_i corresponding to the i^{th} character sequence. We performed the convolution over the entire sequence of characters, i.e., $c_{1:n}$, to obtain a feature map s. We repeated this procedure for L different filter sizes and applied max-pooling [17] to choose the most representative feature corresponding to each filter. Finally, we passed this L dimensional vector through a fully connected softmax layer and projected it over a number of distinct ethnicities. We added dropout [27] as a regularizer.

EthniColr. Inspired by the work of Hofstra et al. [28], we used Ethnicolr[4], the publicly available library from Sood and Laohaprapanon [62], to predict an individual's race/ethnicity from their full name. Ethnicolr employs a neural architecture based on Long Short Term Memory Networks (LSTMs) to model the relationship between the sequence of characters in a name and race/ethnicity.

3.3.2 Homophily-based Inference Algorithms. We chose two algorithms that exploit sociocultural homophily (in addition to other data) to predict peoples' race/ethnicity.

BISG. Prior work in the domain of finance and healthcare [1, 12] has employed the Bayesian Improved Surname Geocoding (BISG) method to predict the race/ethnicity of individuals. This algorithm is based on the observation that racial/ethnic demographics tend to exhibit spatial homophily, i.e., living in similar neighborhoods. We used the publicly available BISG tool[5] in our study, which uses an individual's last name and Zip Code to predict race/ethnicity.

NamePrism. We used the NamePrism API[6] by Ye et al. [66] for race/ethnicity classification. Motivated by the observation that individuals frequently communicated with peers of similar age, language, and location [40], Nameprism exploits the homophily

[3] https://www.ncsbe.gov/results-data/voter-registration-data

[4] https://github.com/appeler/ethnicolr

[5] https://github.com/theonaunheim/surgeo

[6] http://www.name-prism.com/

phenomena in email contact lists to create name embeddings that can be used to predict race/ethnicity.

For the four algorithms in § 3.3.1 and § 3.3.2, we trained a CNN architecture similar to EthCNN, called GenCNN, to predict whether a given name is assigned to the group 'Men' or 'Women'. Since GenCNN achieved a high F1-score of 0.97 for both categories, we assume the inference for gender to be perfect for text-based algorithms and pass the actual gender with the predicted ethnicity for inference.

3.3.3 Facial Analysis-based Inference Algorithms. Finally, we selected one algorithm that relies on facial analysis.

DeepFace. We used the public wrapper [60] for DeepFace by Facebook [63] to obtain DeepFace's error rates when classifying race/ethnicity and gender from the FairFace dataset [33].

4 EXPERIMENTS

In this section, we outline the three experiments that we performed to examine the relationship between inferred demographics and fair ranking.

4.1 Simulations

In our first experiment, we examined the relationship between demographic mis-classification and fair ranking guarantees under controlled conditions by performing simulations using synthetic data. We used a modified version of the synthetic ranked list generation method discussed in Geyik et al. [23], as follows:

(1) We manually crafted five ground-truth probability distributions P for the protected attributes of the simulated people. The distributions, labeled A through E and shown in Table 1, each contained three or four groups. These were the target distributions of our fairly re-ranked lists.
(2) For each probability distribution P, we generated 1,000 people per group $g_i \in P$ and assigned each a random utility score $s_i \in [0, 1]$. We then sorted the combined list of people in decreasing order of s_i to generate the ranking τ.
(3) We ran the *DetConstSort* algorithm discussed in § 3.1 with the desired distribution P and τ as inputs to produce the fairness-aware re-ranked list τ_f. $|\tau_f| = 300$.
(4) We calculated NDKL, ABR, DTBR, and DIBR on τ and τ_f.
(5) We repeated steps 2–4 100 times and computed the mean values for our metrics.
(6) We repeated steps 2–5 for demographic prediction accuracies varying from 0.1 to 1.0. For instance, an accuracy of 0.1 meant that the attribute g_i was predicted correctly 10% of the time and therefore, in our simulation, we mis-classify g_i as any g_j where $j \neq i$ 10% of the time.

Table 1 shows the mean fairness metrics for our empirical distributions before running *DetConstSort*.

4.2 Case Study: COMPAS Dataset

Our second experiment was a case study leveraging the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset. COMPAS is an algorithm developed by NorthPointe Inc. used by judges and parole officers in the U.S. justice system to predict a recidivism score (i.e., probability of reoffending if released) for a criminal defendant.

Distribution	NDKL	ABR
Dist A (W: 0.33, B: 0.33, A: 0.33)	0.08	0.66
Dist B (W: 0.2, B: 0.3, A: 0.5)	0.08	0.71
Dist C (W: 0.1, B: 0.3, A: 0.6)	0.30	0.86
Dist D (W: 0.1, B: 0.2, A: 0.7)	0.37	0.91
Dist E (W: 0.1, B: 0.2, A: 0.6, L: 0.1)	0.40	0.60

Table 1: Fairness metric values computed between the target distribution on the left and a randomly generated unfair distribution. A: Asian, B: Black, L: Latinx, and W: White

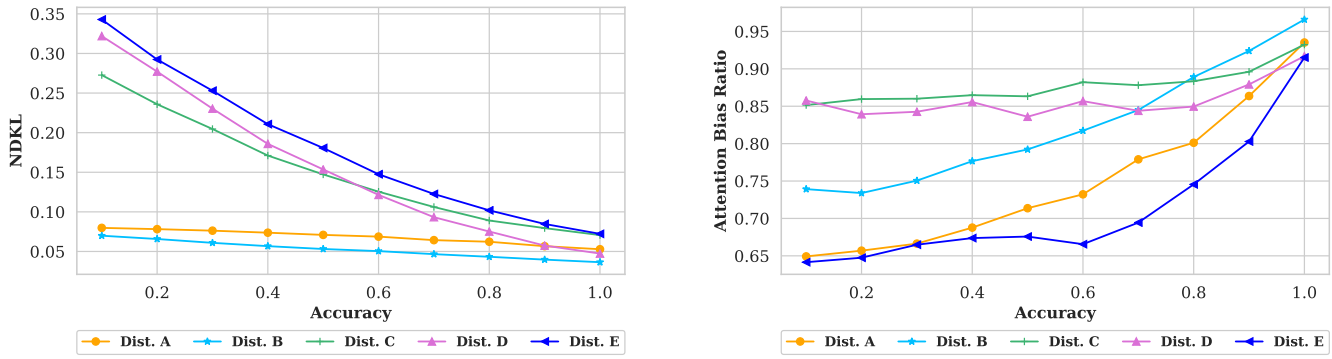
Angwin et al. [3] at ProPublica investigated and published a dataset of the criminal defendants processed through the COMPAS algorithm in Florida's Broward County who had either reoffended within two years or had at least two years outside of a correctional facility. The dataset contains features like age, gender, race/ethnicity, degree of charge, time spent in prison, etc. for 6,172 defendants. ~80% of people in the dataset are men; 51.4% are Black, 34% are White, and 8% are Latinx/Hispanic. The remaining ethnic/racial identities are too small to be statistically significant. Angwin et al. [3] and follow-up studies [67] showed that COMPAS was biased in favor of White defendants at the expense of Black defendants.

Although COMPAS was designed for binary classification, we adapted it for ranking to suite our study. We performed Logistic Regression (LR) using all the features in the dataset to predict whether a defendant would recidivate, then used the output probability of the fitted LR model $\in [0, 1]$ as a score s_i to rank the defendants in descending order, thus producing the baseline ranking τ. The ROC AUC score for the fitted LR model was 0.74.

To analyze fairness, we computed NDKL, ABR, DTBR, and DIBR on τ and τ_f, i.e., the re-ranking produced by *DetConstSort* given the ground-truth demographics of the defendants. We refer to these results as "Baseline" and "Oracle." We also computed our four metrics on re-rankings produced by *DetConstSort* after introducing demographic mis-classifications, based on the error rates we empirically derived from the five demographic inference algorithms we introduced in § 3.3.

4.3 Case Study: Adult Income Dataset

Our third experiment was also a case study, this time leveraging the Adult Income Dataset from the UC Irvine Machine Learning Repository. This data was extracted from the 1994 US Census Bureau database and includes features like age, gender, ethnicity/race, education, occupation, etc. The task for which the dataset was created was to train a classifier to predict whether an adult with the given characteristics made $> 50K$ USD per year. The demographics of the people in the dataset are not balanced: ~66% are men and 85% are white. For the purposes of our study, we retained data related to Whites, Blacks, and Asians; people of other ethnicities/races appeared too infrequently to be analyzed with statistical confidence. The entire dataset contained 48,842 people, of which we randomly subsampled 7,395 (~15%) for our case study.



(a) NDKL

(b) Attention Bias Ratio

Figure 3: Distributions of NDKL and ABR scores for fairly-ranked lists as demographic inference accuracy was varied, based on simulations using synthetic data. For details about the ground-truth population distributions, refer to Table 1.

Although this dataset was also intended for binary classification (i.e., predicting income > $\$50K$ or ≤ $\$50K$), we again adapted it for our purposes. Similar to the COMPAS dataset, we trained and fit a LR model on the dataset, used the probability scores s_i to rank the people in descending order, and produce our baseline ranked list τ. The ROC AUC score for the fitted LR model was 0.88.

We evaluated the ranking fairness of this dataset using the same approach as for the COMPAS dataset.

5 RESULTS

Having discussed our methods and the structure of our experiments, we now present our results. For brevity, we focus on NDKL and ABR as metrics of fairness since the results from DITR and DIBR were almost always in agreement with ABR. We present the results of all four metrics in the Supplementary Materials.

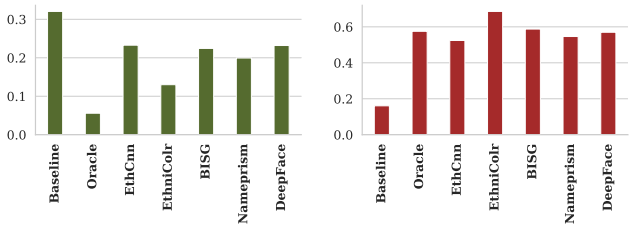
5.1 Simulations

Based on our simulations using synthetic data, we make three observations. *First*, all of the fairness metrics suffer in proportion to the error rate of demographic inference. As shown in Figure 3, NDLK falls (i.e., approaches representational fairness), and ABR rises (i.e., approaches attention parity) as the accuracy of demographic inference increases. This result is intuitive: we cannot expect *DetConstSort* to perform at its best when the underlying demographic data is inaccurate.

Second, we observe varying fair ranking performance with respect to our five ground-truth population distributions. *DetConstSort* was able to achieve low NDKL scores for relatively-uniform distributions, like A and B, regardless of inference accuracy, but struggled to achieve high ABR scores at lower accuracies. Conversely, *DetConstSort* achieves relatively high ABR scores but low NDKL scores for three-group distributions that had an overwhelming majority group, like C and D. Distribution E appears to be a worst-case scenario, combining a clear majority group with three other, much smaller minority groups. These findings demonstrate that there are complex interactions between the composition of the underlying population, accuracy of inference, and fairness guarantees.

Third, by comparing the baseline NDLK and ABR values for non-fairness aware rankings of our five populations from Table 1 to the fairness-aware results in Figure 3, we observe that there are cases where the former has better fairness scores than the latter, depending on the accuracy of demographic inference. This finding shows that the use of a fair ranking algorithm is not categorically better than a non fairness-aware algorithm, depending on the accuracy of the underlying demographic data.

5.2 COMPAS Recidivism Dataset



(a) NDKL

(b) Attention Bias Ratio

Figure 4: NDLK and ABR scores for seven different rankings of defendants from the COMPAS dataset.

Recall that our experimental methodology for our case studies was different than for our simulations. In the case of COMPAS, the dataset contains the ground-truth race and gender of each defendant, as well as their predicted recidivism score. We sorted the defendants by recidivism score and treated it as the baseline, unfair ranking. Using *DetConstSort*, we re-ranked the list using the ground-truth data to produce the "most fair" ranking, which we treated as our oracle. Finally, we intentionally introduced errors into the demographic data based on the error rates we empirically observed from five real-world inference algorithms (see § 3.3), then re-ranked the lists using *DetConstSort*. In this section we refer to intersectional combinations of race/ethnicity and gender as a *group*.

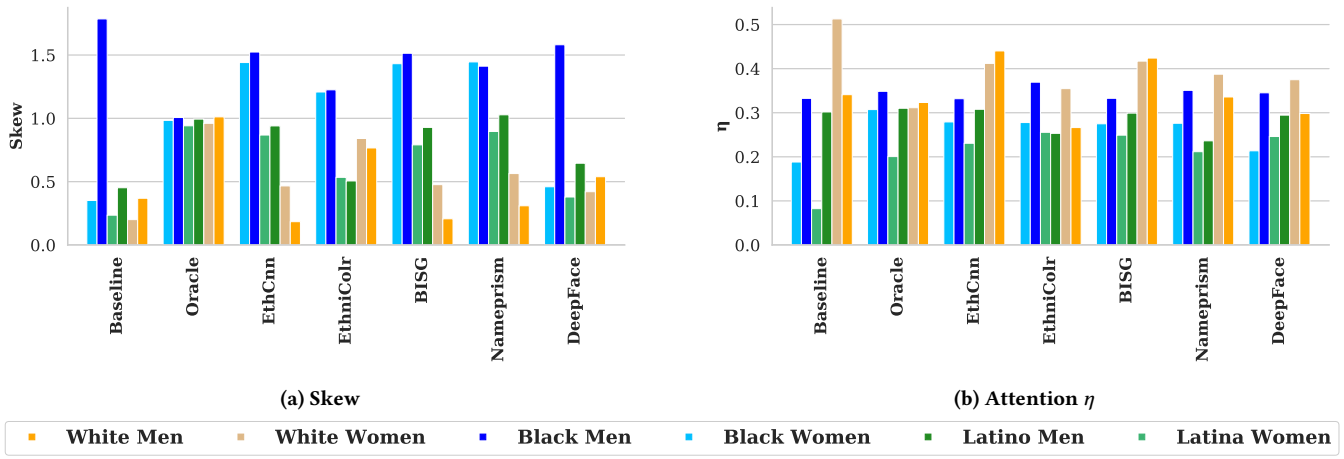


Figure 5: Skew and η scores for individual groups across seven different rankings of defendants from the COMPAS dataset.

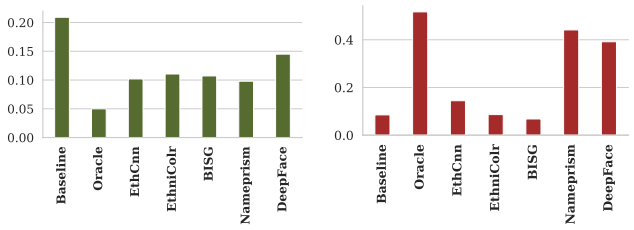
Group	BL	OR	ECLR-P	ECLR-A	CNN-P	CNN-A
Black Men	243	137	9	162	99	207
White Men	30	85	264	69	177	15

Table 2: Frequency counts of Black and White men in the top 300 defendants in the baseline (BL), oracle (O), Ethnicolr (ECLR), and EthCNN (CNN) rankings. -P and -A denote the predicted and actual numbers of defendants in the ranking, i.e., accounting for errors in demographic inference.

Figure 4 shows the overall NDKL and ABR scores for our seven ranked lists of COMPAS defendants. We observe that all of the "fair" rankings outperform the baseline ranking's NDKL score, but the rankings based on inferred demographics all perform worse than the oracle. These results are somewhat surprising, given that *DetConstSort* is optimized for NDKL. In contrast, the differences between the rankings are less pronounced for ABR: the baseline performance is low, but the others perform somewhat comparably — Ethnicolr is even able to outperform the oracle. This odd finding motivated us to probe deeper into the fairness metrics at the group, rather than aggregate-level.

Figure 5a shows the distributions of Skew (the groupwise metric that underlies NDKL) for the groups in each of our seven ranking scenarios. We observe that Black men appear at disproportionately high ranks relative to other groups in the baseline ranking, i.e., they are predicted to have higher rates of recidivism overall, which translates into a high skew. The *DetConstSort* oracle is able to correct this issue, achieving a roughly uniform Skew of 1.0 across all groups. However, we observe that errors in demographic inference cause Skew to become even worse than the baseline scenario: Black men and women become over-represented when relying on inferences from EthCNN, Ethnicolr, BISG, and NamePrism because these algorithms frequently mis-classify Black people as White (e.g., 65% and 79% of Black people are inferred as White by EthCNN and EthniColr, respectively). Table 2 helps explain this issue: although Black men are over-represented in the baseline ranking (243 Black men among the top 300 defendants ranked by recidivism score) demographic mis-classifications create the impression that White men are over-represented (the ECLR-P and CNN-P columns in Table 2), which then causes *DetConstSort* to over-correct, ultimately placing an unfair quantity of Black defendants into the "fair" ranking. Skew for the Deepface ranking deviates from the other inference algorithms because Deepface erroneously predicts a majority of Black women to be Black men, thus increasing the skew of Black men at the expense of Black women.

Figure 5b shows the distributions of mean attention score η (the groupwise metric that underlies ABR) for the groups in each of our seven ranking scenarios. We observe that η scores are disproportionately high and low for White women and Latina women, respectively, in the baseline ranking. Three out of five White women in the dataset occupied the 18^{th}, 53^{rd}, and the 60^{th} position in the baseline ranking, resulting in relatively high attention across a small group. Likewise, there was only one Latina woman in the dataset and she appeared at the 192^{nd} rank, resulting in a low attention score for that group. We observe that *DetConstSort* was able to mostly mitigate these issues and distribute attention evenly when producing the oracle ranking. However, we see that the situation worsened again when we consider errors in demographic inference: since the inference algorithms tended to predict that the original ranked list had a large number of White men (i.e., erroneously classified Black men), the *DetConstSort* algorithm responded by moving White men with high recidivism scores to the top of the ranking (increasing their η scores) before populating the lower ranks with defendants from other groups. Note that we do not observe this general trend in the EthniColr ranking since, in this scenario, the predicted Black men with low recidivism scores at the bottom of the "fair" ranked list were actually White men, thereby lowering η for White men.



(a) NDKL (b) Attention Bias Ratio

Figure 6: NDLK and ABR scores for seven different rankings of people from the Adult Income dataset.

5.3 Adult Income Dataset

We performed a similar set of analyses on the Adult Income Dataset as the COMPAS dataset, to observe the impact of imperfect inference on the fairness ranking metrics. Beginning with the aggregate fairness results depicted in Figure 6, we see that the baseline ranking is relatively unfair according to both the NDKL and ABR metrics, while the oracle ranking is fairest (or at least as fair) as all the other rankings. Among the rankings that incorporated errors in demographic inference, DeepFace yields the least fair ranking according to the NDKL metric, while EthCNN, EthniColr, and BISG yield the most unfair rankings according to the ABR metric.

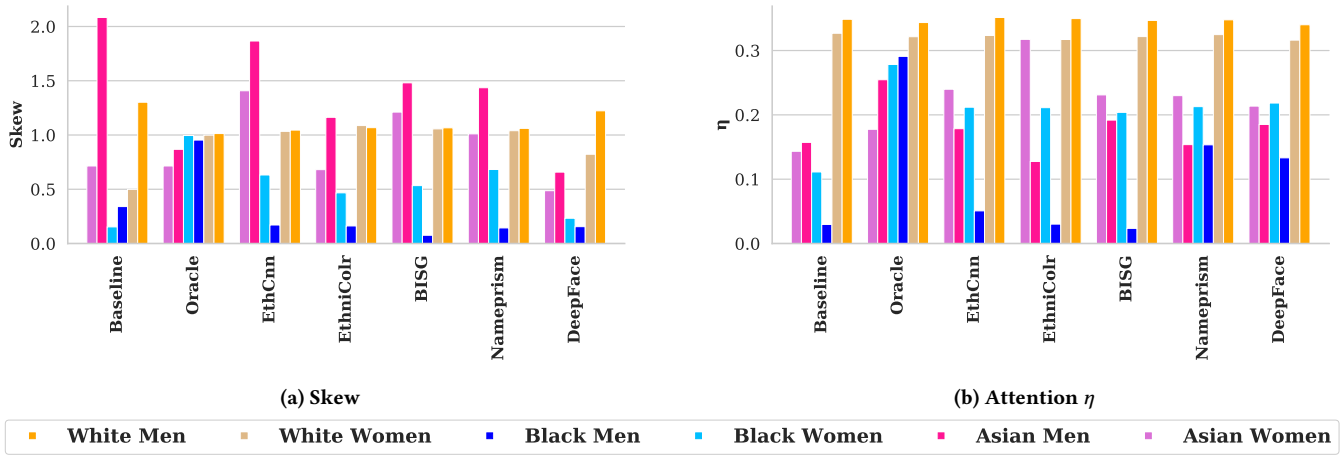
We probe further into our results by inspecting the group-wise values of the fairness metrics, as shown in Figure 7a. Focusing on the baseline ranking, we see that Asians and White men are disproportionately represented amongst the top 300 ranked candidates by their high values of Skew. We note, however, that Asian men constitute only 12 slots, while the lion's share of the ranked list is comprised of White men (239 people). *DetConstSort* thus attempts to minimize Skew in the various "fair" rankings by increasing the representation of other groups in lieu of Asians and White men. In the oracle scenario with ground-truth demographics *DetConstSort* largely succeeds at equalizing Skew across the groups, but not in the cases with erroneous demographic inferences. For example, Ethnicolr incorrectly identifies ten Black men in the dataset when there are only five; as a result *DetConstSort* ends up under-sampling true positive Black men, ultimately producing a top 300 ranking that only includes two Black men. Conversely, because White men are over-represented in this dataset, inference errors barely impact their standing in the rankings.

We observe that demographic mis-classification reinforces the problem of low attention scores for minorities. As shown in Figure 7b, in the baseline ranking Whites dominate the top ranks of the list. Armed with ground-truth demographic information, *DetConstSort* is able to mitigate this disparity to some extent, but not completely because (1) there are so few minorities in the dataset overall and (2) Whites have higher income predictions on average. As with Skew, we find that *DetConstSort* struggles to equalize the η scores across groups when inferred demographic data is used, especially with respect to Black men. Asians and Black women mostly receive more attention after re-ranking relative to the unfair baseline, but not nearly enough to close the gap with Whites.

In summary, these results present a "rich get richer" scenario where (1) pre-existing social and financial inequalities that benefit Whites are compounded by (2) demographic classification algorithms that are more accurate for Whites, thus solidifying the dominance of Whites in the rankings. In real-world banking scenarios, this confluence of factors could potentially create situations where minorities are denied access to banking services or credit, i.e., digital redlining.

6 DISCUSSION

In this study, we presented three experiments that delved into the interactions between demographic inference algorithms and fair ranking algorithms. Using the *DetConstSort* ranking algorithm as a representative example, we examined whether it was able to achieve four fairness benchmarks when faced with erroneous demographic inferences drawn from five real-world algorithms. To ensure realism and comprehensiveness in our experiments, we derived the error rates for the demographic inference algorithms from real-world, ground-truth datasets, and presented results from



(a) Skew (b) Attention η

Figure 7: Skew and η scores for individual groups across seven different rankings of people from the Adult Income Dataset.

controlled simulations and real-world case studies. The takeaway from our experiments is that using inferred demographic data as input to fair ranking algorithms can invalidate their fairness guarantees in ways that are (1) difficult to predict and (2) often harm vulnerable groups of people.

It would have been a positive, pragmatic result if our study found that fair ranking under uncertain inference, even with its limitations, was categorically fairer than non-fairness-aware ranking. Unfortunately, this is not the case: across a range of scenarios, we observe complex interactions between errors in demographic inference and demographic representation in ranked lists. In some cases, groups that were not disadvantaged in the baseline, non-fairness aware ranking become disadvantaged under fair ranking due to errors in inference (e.g., Black women in the COMPAS case study). This comes in addition to existing inequities in the baseline ranking that may not be remedied in the "fair" ranking due to bad inferences. In short, uncertain inference can make bad ranking situations even worse.

The primary takeaway from our study is that relying on inferred demographic data as input for fairness-aware algorithms is dangerous. We could cynically summarize this situation as an instance of the well-known garbage-in/garbage-out principle. However, this oversimplifies the issue in several ways:

(1) Inferred demographic data papers-over fundamental problems in system design, such as a failure to deeply engage with the socio-technical complexities of a given context and collect appropriate data to mitigate biases.
(2) When used in combination with fairness-aware algorithms, inferred demographic data creates a false sense of confidence that underlying biases in data have been mitigated, when in fact the fairness guarantees may be violated in ways that are difficult to predict or bound.
(3) Inferred demographic data permits system designers to truthfully claim that they have adopted technical measures to "de-bias" their data, even though these measures are illusory.

6.1 Uncertainty-aware Ranking

One tempting solution to the problem at hand is uncertainty-aware fair ranking algorithms [15] — if demographic uncertainty could be bounded, then the fair ranking algorithm could be modified to take this into account by adjusting the probabilities associated with each individual, thus producing fair rankings in expectation.

In practice, there are several shortcomings with this idea. *First*, the addition of uncertainty means that the ranking algorithm will not be able to guarantee that any given realization of results is fair. Instead, the ranker will only be able to achieve fairness on average, over the realization of many rankings. There may be real-world situations when this is acceptable: for example, on a dating website, each user can be expected to make many searches for partners over time, so each user will eventually observe a demographically fair set of partners. In other situations fairness-on-average is not acceptable: for example, on a resume search website, a recruiter may only search a handful of times to collect candidates for a specific job. Thus, although the system may present fair results on average, the set of job candidates presented for any given career opportunity will not be demographically fair.

Second, this discussion is predicated on the ability to bound the error rates from a given demographic inference algorithm, which may not be feasible in practice. Even if the error rate can be measured for each demographic group of interest, the uncertainty-aware ranking algorithm must adopt the worst-case error rate from among the groups. This means that the convergence time for fairness-on-average is intrinsically linked to the group with the greatest uncertainty.

6.2 The Fundamental Limits of Inference

Another tempting solution to the issues we have raised is to dedicate more resources towards improving the predictive performance of demographic inference algorithms. However, even if we were to somehow "fix" the inaccuracies of existing demographic inference algorithms, using inferred demographics to achieve algorithmic fairness is normatively problematic in a number of ways that belie simple, technical solutions.

First, these algorithms rob people of their autonomy by placing them into groups defined by the algorithm developer, rather than allowing people to define themselves on their own terms.

Second, the groups that are commonly "supported" by existing inference algorithms reify problematic categories, like sex and gender binaries or politically constructed racial categories. These problems are evident in our study: for example, in our case studies we are limited to examining fairness for four racial and ethnic categories that (1) are disputed and based on a history of institutional racism, (2) are not inclusive of many marginalized groups, and (3) fail to grapple with intersectional groups and multiracial people. Similar issues are evident with respect to gender, i.e., we are limited to examining fairness for binary genders, whereas comprehensive approaches would disambiguate sex from gender and treat each as multifaceted spectra.

Third, even if demographic inference algorithms were redesigned for inclusivity, they are still fundamentally retrospective and fixed. At best, demographic inference algorithms will always draw on (harmful) stereotypes for classification. At worst, they are no better than digital phrenology or physiognomy [44], relying on spurious correlations to construct false archetypes.

6.3 Norms and Policy

The results of our study highlight once again the tension between norms and policy on one side, and the desire for fair sociotechnical systems on the other. To achieve demographically fair algorithms in practice, there is simply no substitute for accurate data about peoples' characteristics. That said, there are many contexts in which people may be reluctant to divulge this information (e.g., any situation involving access to opportunities like employment and housing) or service providers are legally forbidden from collecting the necessary data (e.g., banking, or situations involving children). Revising the law to permit demographic data collection, provided that it will only be used to ensure fairness, may be a necessary first step, but even then it will take time and concerted effort to convince people that divulging this information is in their best interest.

In the meantime, the results of our study demonstrate that attempting to rely on inferred demographic data to sidestep these issues is challenging at best, and folly at worst.

REFERENCES

[1] Dzifa Adjaye-Gbewonyo, Robert A Bednarczyk, Robert L Davis, and Saad B Omer. 2014. Using the Bayesian Improved Surname Geocoding Method (BISG) to create a working classification of race and ethnicity in a diverse managed care population: a validation study. *Health services research* 49, 1 (2014), 268–283.

[2] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. 2019. Fair regression: Quantitative definitions and reduction-based algorithms. *arXiv preprint arXiv:1905.12843* (2019).

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2019. Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. 2016. *URL https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing* (2019).

[4] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *104 California Law Review* 671 (2016).

[5] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943* (2018).

[6] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409* (2017).

[7] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in Recommendation Ranking through Pairwise Comparisons. In *KDD*. https://arxiv.org/pdf/1903.00780.pdf

[8] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *The 41st international acm sigir conference on research & development in information retrieval*. 405–414.

[9] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*. 4349–4357.

[10] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *International Conference on Machine Learning*. 803–811.

[11] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.

[12] Consumer Financial Protection Bureau. 2014. Using publicly available information to proxy for unidentified race and ethnicity. *Report available at http://files. consumerfinance. gov/f/201409_cfpb_report_ proxy-methodology. pdf* (2014).

[13] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. 2019. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 319–328.

[14] L Elisa Celis, Lingxiao Huang, and Nisheeth K Vishnoi. 2020. Fair Classification with Noisy Protected Attributes. *arXiv preprint arXiv:2006.04778* (2020).

[15] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. 2018. Ranking with Fairness Constraints. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[16] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the Impact of Gender on Rank in Resume Search Engines.

[17] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research* 12, Aug (2011), 2493–2537.

[18] Nicholas Diakopoulos, Daniel Trielli, Jennifer Stark, and Sean Mussenden. 2018. I Vote For—How Search Informs Our Choice of Candidate. In *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple*, M. Moore and D. Tambini (Eds.). 22.

[19] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[20] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. *Berkman Klein Center Research Publication* 2020, 1 (2020). https://ssrn.com/abstract=3518482

[21] Joel Escudé Font and Marta R Costa-Jussa. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116* (2019).

[22] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*. 329–338.

[23] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2221–2231.

[24] Naman Goel, Mohammad Yaghini, and Boi Faltings. 2018. Non-discriminatory machine learning through convex fairness criteria. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 116–116.

[25] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr.

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

[27] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* (2012).

[28] Bas Hofstra, Vivek V Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A McFarland. 2020. The Diversity–Innovation Paradox in Science. *Proceedings of the National Academy of Sciences* 117, 17 (2020), 9284–9291.

[29] Lingxiao Huang and Nisheeth K Vishnoi. 2019. Stable and fair classification. *arXiv preprint arXiv:1902.07823* (2019).

[30] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. 2020. Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1 (May 2020).

[31] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2017. Fairness in reinforcement learning. In *International Conference on Machine Learning*. PMLR, 1617–1626.

[32] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.

[33] Kimmo Kärkkäinen and Jungseock Joo. 2019. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913* (2019).

[34] Anna Kawakami, Khonzoda Umarova, Dongchen Huang, and Eni Mustafaraj. 2020. The 'Fairness Doctrine' Lives on? Theorizing about the Algorithmic News Curation of Google's Top Stories. In *Proc. of HT*.

[35] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations.

[36] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1746–1751. https://doi.org/10.3115/v1/D14-1181

[37] Michal W. Kosinski and Yilun Wang. [n.d.]. Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images. *Journal of Personality and Social Psychology* 114, 2 (Feb. [n. d.]), 246–257.

[38] Caitlin Kuhlman, MaryAnn VanValkenburg, and Elke Rundensteiner. 2019. Fare: Diagnostics for fair ranking using pairwise error metrics. In *The World Wide Web Conference*. 2936–2942.

[39] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi, and Karrie Karahalios. 2017. Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media.

[40] Jure Leskovec and Eric Horvitz. 2008. Planetary-Scale Views on a Large Instant-Messaging Network. In *Proceedings of the 17th International Conference on World Wide Web* (Beijing, China) *(WWW '08)*. Association for Computing Machinery, New York, NY, USA, 915–924. https://doi.org/10.1145/1367497.1367620

[41] Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. 2018. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859* (2018).

[42] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[43] Emma Lurie and Eni Mustafaraj. 2018. Investigating the Effects of Google's Search Engine Result Page in Evaluating the Credibility of Online News Sources.

[44] Greggor Mattson. 2017. Artificial Intelligence Discovers Gayface. Sigh. Greggor Mattson Personal Blog. https://greggormattson.com/2017/09/09/artificial-intelligence-discovers-gayface/.

[45] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[46] Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*. 107–118.

[47] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. 2020. Controlling Fairness and Bias in Dynamic Learning-to-Rank. *arXiv preprint arXiv:2005.14713* (2020).

[48] Ankan Mullick, Sayan Ghosh, Ritam Dutt, Avijit Ghosh, and Abhijnan Chakraborty. 2019. Public Sphere 2.0: Targeted Commenting in Online News Media. In *European Conference on Information Retrieval*. Springer, 180–187.

[49] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, Vol. 2018. NIH Public Access, 1931.

[50] Jakob Nielsen. 2003. Usability 101: introduction to usability. Jakob Nielsen's Alertbox.

[51] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366 (Oct. 2019).

[52] Osonde A Osoba and William Welser IV. 2017. *An intelligence in our image: The risks of bias and errors in artificial intelligence.* Rand Corporation.

[53] Amifa Raj, Connor Wood, Ananda Montoly, and Michael D Ekstrand. 2020. Comparing Fair Ranking Metrics. *arXiv preprint arXiv:2009.01311* (2020).

[54] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[55] Ronald E Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM: Human-Computer Interaction* 2, CSCW (November 2018).

[56] Ronald E. Robertson, Shan Jiang, David Lazer, and Christo Wilson. 2019. Auditing Autocomplete: Recursive Algorithm Interrogation and Suggestion Networks.

[57] Ronald E Robertson, David Lazer, and Christo Wilson. 2018. Auditing the Personalization and Composition of Politically-Related Search Engine Results Pages.

[58] José A Sáez, Mikel Galar, JuliáN Luengo, and Francisco Herrera. 2013. Tackling the problem of classification with noisy data using multiple classifier systems: Analysis of the performance and robustness. *Information Sciences* 247 (2013), 1–20.

[59] Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the Impact of User Attentionon Fair Group Representation in Ranked Lists. In *Companion Proceedings of The 2019 World Wide Web Conference*. 553–562.

[60] Sefik Ilkin Serengil and Alper Ozpinar. 2020. LightFace: A Hybrid Deep Face Recognition Framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE.

[61] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2219–2228.

[62] Gaurav Sood and Suriyan Laohaprapanon. 2018. Predicting Race and Ethnicity From the Sequence of Characters in a Name. arXiv:1805.02109 [stat.AP]

[63] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1701–1708.

[64] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.

[65] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. 1–6.

[66] Junting Ye, Shuchu Han, Yifan Hu, Baris Coskun, Meizhu Liu, Hong Qin, and Steven Skiena. 2017. Nationality Classification Using Name Embeddings. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (Singapore, Singapore) *(CIKM '17)*. Association for Computing Machinery, New York, NY, USA, 1897–1906. https://doi.org/10.1145/3132847.3133008

[67] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) *(WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1171–1180. https://doi.org/10.1145/3038912.3052660

[68] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*. PMLR, 962–970.

[69] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1569–1578.

[70] Meike Zehlike and Carlos Castillo. 2020. Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of The Web Conference 2020*. 2849–2855.

SUPPLEMENTARY MATERIAL

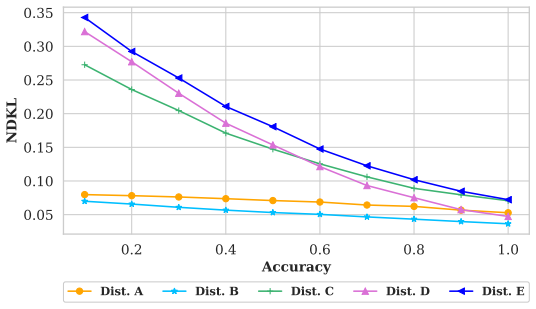
Simulation results In this section we note the relationship between the fairness metrics and the accuracy of the inference algorithm. At the outset, we observe that the fairness metrics suffer directly proportionally to the error rate of demographic inference in Figure 8. As the accuracy increases, NDKL falls (i.e., achieves representational fairness), ABR rises (i.e., approaches attention parity), DTBR rises (i.e., approaches disparate treatment parity), and DIBR rises (i.e., approaches disparate impact parity). We also observe that DTBR in Figure 8c and DIBR in Figure 8c follow a similar trend to ABR in Figure 8b. Since the underlying scores for different groups in the simulation are generated at random from a uniform distribution, the bias ratios for the different groups cancel out on average and hence the DTBR and DIBR ratios that we obtain are proportional to ABR.

Distribution	NDKL	ABR	DTBR	DIBR
Dist A (W: 0.33, B: 0.33, A: 0.33)	0.08	0.66	0.67	0.65
Dist B (W: 0.2, B: 0.3, A: 0.5)	0.08	0.71	0.71	0.70
Dist C (W: 0.1, B: 0.3, A: 0.6)	0.30	0.86	0.86	0.86
Dist D (W: 0.1, B: 0.2, A: 0.7)	0.37	0.91	0.91	0.91
Dist E (W: 0.1, B: 0.2, A: 0.6, L: 0.1)	0.40	0.60	0.60	0.59

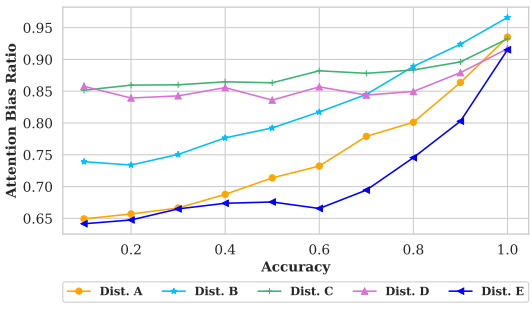
Table 3: Fairness metric values computed between the target distribution on the left and a randomly generated unfair distribution. A: Asian, B: Black, L: Latinx, and W: White

COMPAS dataset We build upon the findings mentioned in § 5.2 for the COMPAS Recidivism dataset. We justify the high attention η in Figure 10b for White men since they were promoted by the ranking algorithm to the top of the list by imperfect demographic inference. However, EthniColr showed a deviation from this trend since these algorithms had erroneously placed White Men with low scores at the bottom of the list, believing them to belong to other groups. This brings down the attention score η for White Men. The scores attributed to these lower-ranked White men are also poor since they were originally at end of the list. This results in a disproportionately low value of γ for White Men for EthniColr as opposed to EthCNN (say) in Figure 10d. Likewise, we observe a higher value of θ for White Men in EthniColr as opposed to EthCNN. Such a finding also helps explain the high value of ABR and DIBR for EthniColr, sometimes even surpassing the Oracle. Although the high values of ABR and DIBR look good on paper, the premise by which the values were obtained is problematic. Due to imperfect demographic inference, several White Men with low scores were promoted to the top of the list since they were inferred to be Black Men or Latinos, in place of White Men with high scores. Thus, such an algorithm violates the very notion of individual fairness.

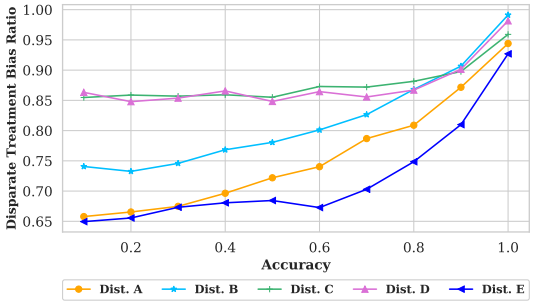
Adult Income Dataset Similarly, we elaborate on the analysis of imperfect demographic inference on the Adult Income Dataset. We observe a similar trend for all metrics, White Men and Women who are disproportionately present in the dataset as well as have the least errors in mis-classification have disproportionate high values of η, θ, and γ, signifying the high skewed values for attention, disparate treatment and disparate impact respectively.



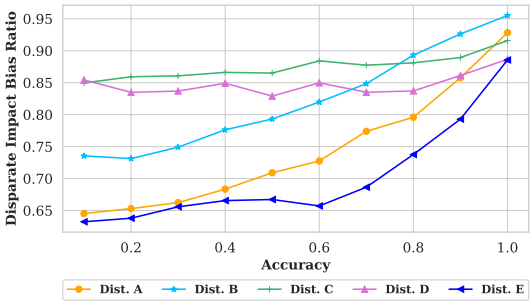
(a) NDKL



(b) Attention Bias Ratio

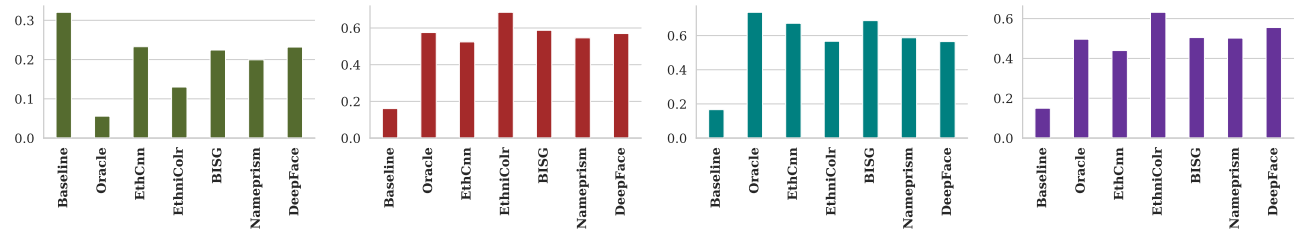


(c) Disparate Treatment Bias Ratio



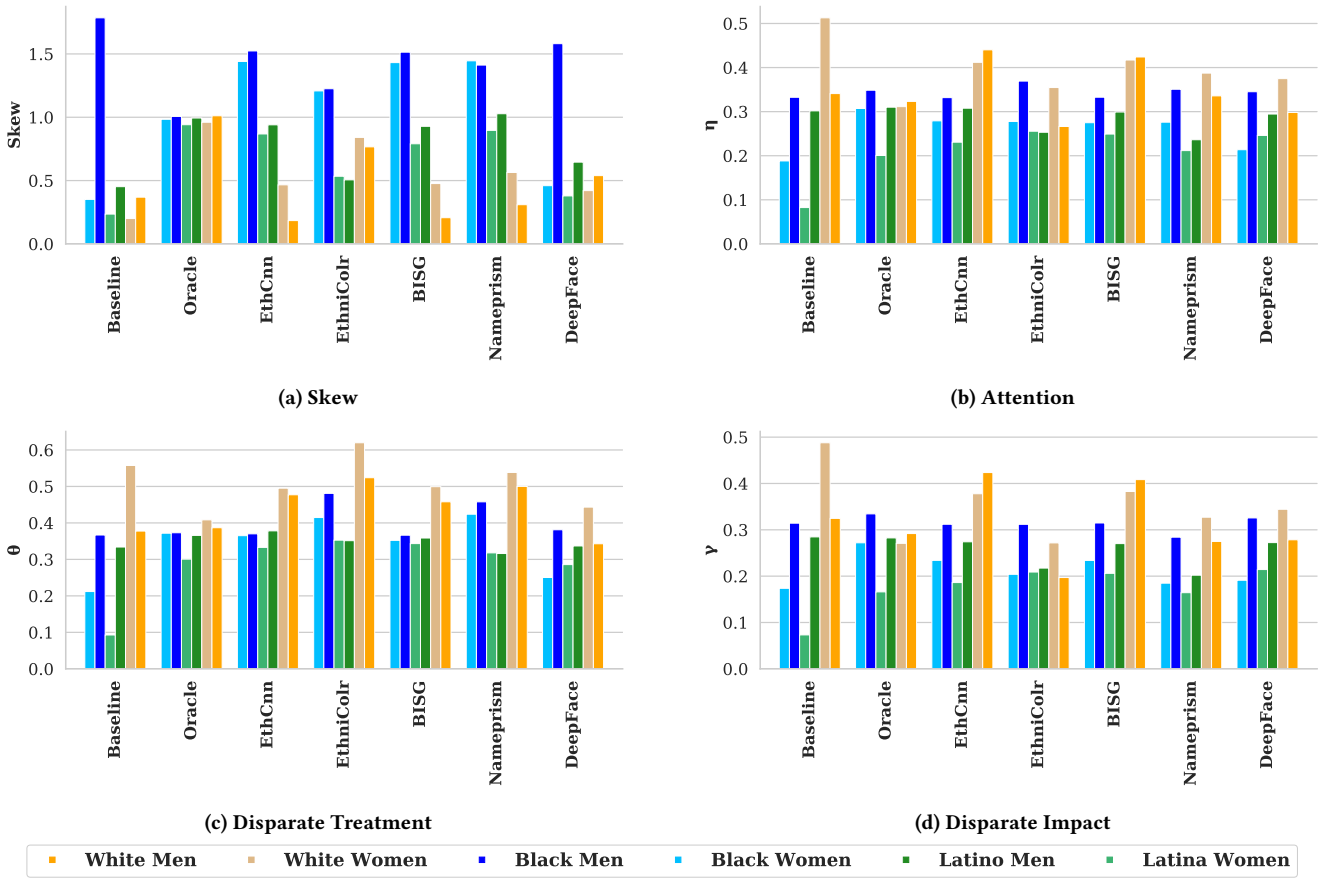
(d) Disparate Impact Bias Ratio

Figure 8: Distributions of NDKL,ABR,DTBR and DIBR scores for fairly-ranked lists as demographic inference accuracy was varied, based on simulations using synthetic data. For details about the ground-truth population distributions, refer to Table 1.



(a) NDKL (b) Attention Bias Ratio (c) Disparate Treatment Bias Ratio (d) Disparate Impact Bias Ratio

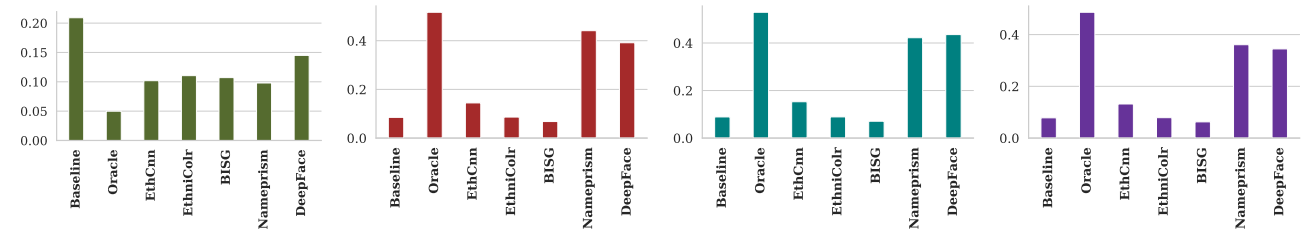
Figure 9: Overall scores for seven different rankings of defendants from the COMPAS dataset.



(a) Skew (b) Attention

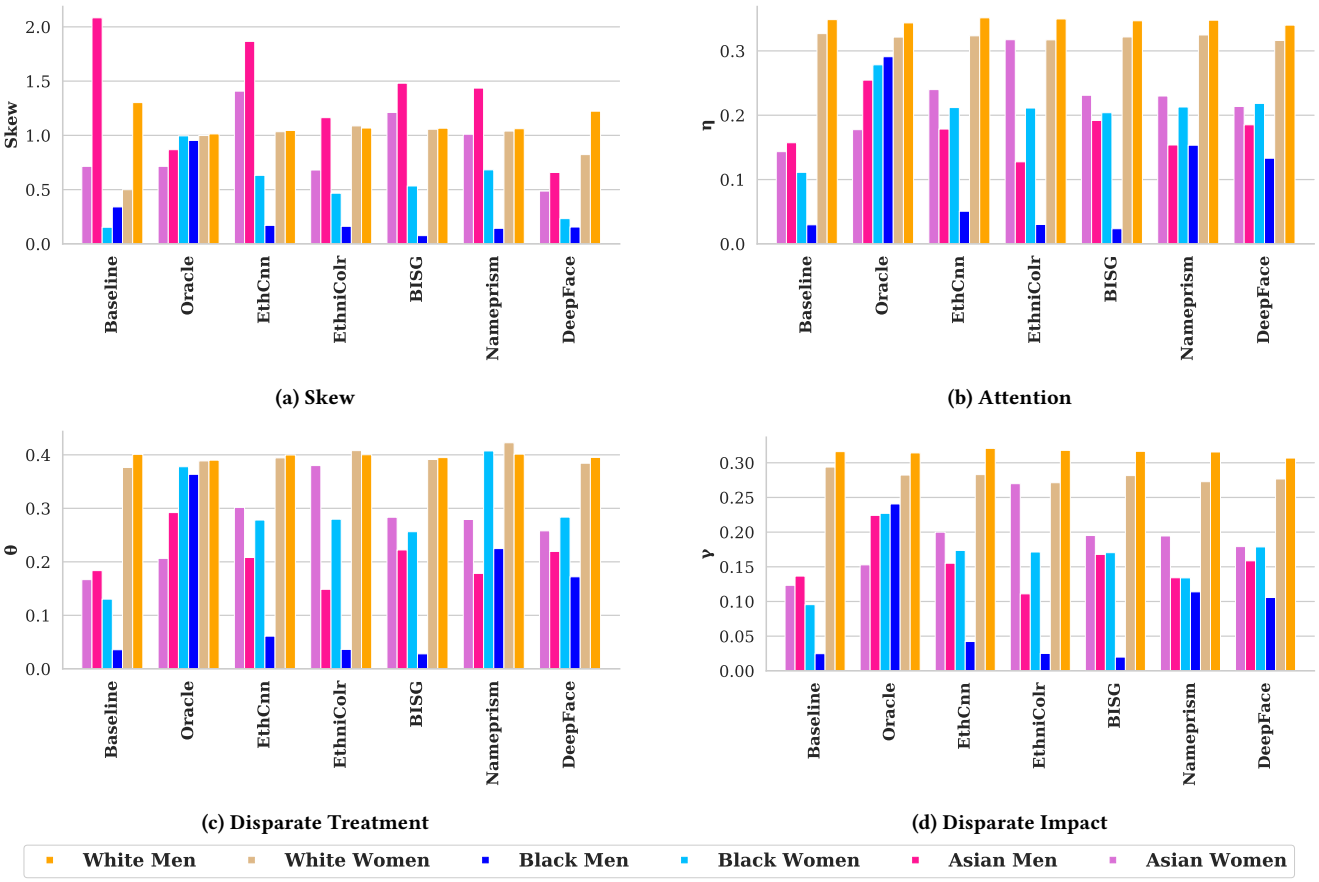
(c) Disparate Treatment (d) Disparate Impact

Figure 10: Skew, η, θ and γ scores for individual groups across seven different rankings of defendants from the COMPAS dataset.



(a) NDKL (b) Attention Bias Ratio (c) Disparate Treatment Bias Ratio (d) Disparate Impact Bias Ratio

Figure 11: Overall scores for seven different rankings of people from the Adult Income dataset.



(a) Skew (b) Attention

(c) Disparate Treatment (d) Disparate Impact

Figure 12: Skew, η, θ and γ scores for individual groups across seven different rankings of people from the Adult Income Dataset.