Hidden or Inferred: Fair Learning-To-Rank with Unknown Demographics

Oluseun Olulana[1], **Kathleen Cachel**[1], **Fabricio Murai**[1,2], **Elke Rundensteiner**[1,2]

[1]Data Science Program, Worcester Polytechnic Institute, Worcester, MA,
[2]Computer Science Department, Worcester Polytechnic Institute, Worcester, MA
{omolulana, kcachel, fmurai, rundenst}@wpi.edu

Abstract

As learning-to-rank models are increasingly deployed for decision-making in areas with profound life implications, the FairML community has been developing fair learning-to-rank (LTR) models. These models rely on the availability of sensitive demographic features such as race or sex. However, in practice, regulatory obstacles and privacy concerns protect this data from collection and use. As a result, practitioners may either need to promote fairness despite the absence of these features or turn to demographic inference tools to attempt to *infer* them. Given that these tools are fallible, this paper aims to further understand how errors in demographic inference impact the fairness performance of popular fair LTR strategies. In which cases would it be better to keep such demographic attributes *hidden* from models versus *infer* them? We examine a spectrum of fair LTR strategies ranging from fair LTR with and without demographic features hidden versus inferred to fairness-unaware LTR followed by fair re-ranking. We conduct a controlled empirical investigation modeling different levels of inference errors by systematically perturbing the inferred sensitive attribute. We also perform three case studies with real-world datasets and popular open-source inference methods. Our findings reveal that as inference noise grows, LTR-based methods that incorporate fairness considerations into the learning process may increase bias. In contrast, fair re-ranking strategies are more robust to inference errors. All source code, data, and experimental artifacts of our experimental study are available here: *https://github.com/sewen007/hoiltr.git*

1 Introduction

Background: Fairness of LTR-based Ranking. Ranked search results are increasingly at the heart of artificial intelligence and automated decision-making systems. As such systems progressively impact our daily lives, there is a growing need to ensure that these technologies do not disproportionately harm or replicate societal biases toward disadvantaged populations or legally protected groups. To this end, the fair machine learning community has developed various fair LTR models (Zehlike and Castillo 2020; Wang, Tao, and Fang 2022) and metrics (Patro et al. 2022; Ekstrand et al. 2021) for assessing such models. At a high level, LTR models learn a scoring function, so that when deployed, the model's learned relevance scores produce an ordering of candidate items. While conventional fairness-unaware LTR methods aim to produce a utility maximizing ordering of candidates, *fairness-aware LTR* methods aim to ensure this ranking is also a *fair* ordering of candidates (items).

Challenges: Fairness without Protected Attributes. Even with such progress, practical obstacles prevent the widespread adoption of these bias mitigation methods. One challenge is that fair LTR models as well as other fairness-enhanced methods, such as fair classifiers, require demographic information associated with candidate items during both model training and subsequent model deployment for real-world use (Dwork et al. 2012). However, in practice, it may not be possible to collect, gain access, or use protected demographic features due to privacy concerns or legal restrictions. For instance, the European Union's GDPR (General Data Protection Regulation) legislation strictly regulates collection, retention, and use of demographic data for algorithmic purposes (Council of the European Union 2016).

Also, there is policy tension between mandates *to ensure algorithms are fair* and mandates *prohibiting the use of demographic data*. One example is the United States credit industry (eCFR 2024). Consumer-facing lenders are largely prohibited from collecting sensitive demographic data; yet United States federal law explicitly prohibits creditors from discriminating on the basis of certain protected demographics and thus leaves them wanting to verify that they did not do so – the later of which requires access to the same sensitive information (Bogen, Rieke, and Ahmed 2020). Thus, companies and institutions are increasingly caught in the middle. Surveys of data scientists and algorithm developers highlight the challenge these practitioners face in promoting fairness without such sensitive data (Holstein et al. 2018).

State-of-Art and their Limitations. A common workaround is to infer demographic data from other available information, such as first names, social media (Fink, Kopecky, and Morawski 2012), or email content (Cheng et al. 2009). However, the accuracy of these estimation (inference) tools can vary significantly (Cheng et al. 2009; Santamaría and Mihaljević 2018).

Recent work in the fair-ranking community has developed fair ranking metrics (Ghazimatin et al. 2022; Kırnap et al. 2021) that account for error rates of demographic attributes. Ghosh, Dutt, and Wilson (2021) investigate the De-

terministic Constrained Sorting (DetConstSort) re-ranking method coupled with the use of demographic inference for integrating fairness. DetConstSort is a fair re-ranking algorithm used as post-processing step. Their study reveals that DetConstSort performs poorly when given inaccurate demographic information. This raises questions regarding the performance of fair LTR strategies that instead choose to either infer or ignore the demographic information. Fair LTR methods have been shown to achieve better fairness-relevance trade-offs compared to applying fair re-ranking methods like DetConstSort to existing fairness-unaware LTR methods – while assuming full access to protected (demographic) attributes (Zehlike and Castillo 2020). This makes these approaches desirable in real-world settings. Therefore, with significance to practitioners, in this work, we ask: *How do errors in demographic inference impact the fairness performance of different fair LTR strategies? Also, how do these errors impact utility?*

Our Approach. We tackle these research questions by investigating the performance of popular fair LTR strategies when deployed in situations where demographic data is unavailable and thus needs to be inferred. We investigate both fair LTR and fair re-ranking type models, as they (and their combinations) cover the majority of fairness-enhanced rank-based machine learning pipelines. We also investigate the case where protected attributes remain hidden as a fairness strategy, meaning, all candidates have the "unknown" as demographic attribute. We study real-world scenarios where models are first trained on available data, with and without sensitive demographic attributes. However, later, during model deployment, data issues may arise that saddle practitioners with a choice to make – to work with data while the sensitive attribute remains *hidden* or to augment the data by *inferring* the missing demographic attributes. While our experiments focus on gender, our study methodology is equally applicable to other protected features.

Our experimental evaluation is composed of two studies (Sections 4.4 and 4.5). In the first study, we systematically perturb the inferred protected attribute to model different levels of inference error under three scenarios. Each scenario is designed to capture the different kinds and levels of errors possible in a real-world scenario. In the second study, we make use of popular demographic inference services (Gender-API n.d.; Namesor n.d.; BehindTheName 1996) to compare and contrast their impact in the context of three real-world data sets. We explore the effect on fairness with respect to ranked candidates' true group identities.

Our investigation has led to the following findings:

- Re-ranking fair strategies that enforce group fairness based on the inferred distribution of test candidates are more robust to inaccurate inference of demographic attributes than fair LTR strategies. This leads to the guidance that, under noisy inference scenarios, practitioners may achieve a higher level of fairness if adopting a fair re-ranking instead of a fair LTR strategy.
- While fairness-aware strategies achieve considerable fairness even when working with inaccurately inferred demographic attributes, fairness decreases as inference errors increase. This suggests that a practitioner's fairness goals may be adversely affected by the utilization of lower-quality inference tools. Practitioners are urged to proceed with caution and verify the quality of their inference tools beforehand.
- In a scenario where demographic attributes are missing, it is better to utilize a fairness-aware model that relies on inferred missing attributes than to adopt a fairness-unaware model that ignores the missing attributes as long as inference errors are relatively low (up to 10%).
- Lastly, we observe that across all three real-world data set case studies, the fairness-unaware models show increased levels of fairness when working with demographic attributes that had been incorrectly inferred. We attribute this phenomenon to candidate items being wrongly classified as members of the alternate group.

2 Related Work

While many ethical considerations in designing ranking systems exist, fairness is typically conceptualized as either individual fairness or group fairness. Individual fairness ensures that similar individuals receive similar outcomes (Dwork et al. 2012). Group fairness ensures that protected groups of people (such as race or gender) receive comparable shares of the positive outcome (Li et al. 2021; Ekstrand et al. 2021). In this work, we consider group fairness, the primary concern of the fair ranking literature (Patro et al. 2022), which aligns with the existing focus of AI regulation (Jillson 2021).

Fairness can be incorporated into an LTR model by adding a fairness parameter or constraint to the learning objective (Zehlike and Castillo 2020; Singh and Joachims 2019; Wang, Tao, and Fang 2022). Alternatively, instead of modifying a fairness-unaware LTR algorithm, fairness can be introduced by reordering a ranking generated by a model (Geyik, Ambler, and Kenthapadi 2019; Zehlike et al. 2017). When striving to be demographically fair, not only is the fairness affected, but also the utility (relevance) of the ranking. In other words, integrating fairness into a ranking framework induces a fairness-utility trade-off. This trade-off has been addressed in some existing work on fairness (Zehlike and Castillo 2020; Li et al. 2022). Zehlike and Castillo (2020), in their fairness-aware LTR loss function, introduce a parameter γ that balances the trade-off between utility and disparate exposure. (Li et al. 2022) emphasize the importance for researchers to explore the relationship between fairness and utility to motivate practitioners to promote fairness. In an earlier paper by (Li et al. 2021), a constrained optimization problem was formulated, where the overall recommendation quality (utility) is the objective function and an upper bound ϵ on group (un-)fairness is enforced via a constraint. The overarching assumption underlying fairness-enhanced algorithms is that demographic data is readily accessible and correct for use by these algorithms (Holstein et al. 2018). Recent work has begun to relax this assumption by developing methods that account for error rates of demographic attributes when incorporating fairness (Mehrotra and Vishnoi 2022; Celis et al. 2021; Wang et al. 2020; Mozannar, Ohannessian, and Sre-

Symbol	Definition
C	List of items $\{c_1, ..., c_n\}$ to be ranked
x_i	Attribute score vector for c_i
τ	A ranking ordering (top is better)
$s(c_i)$	Ground-truth score of item $c_i \in C$
$s_\tau(j)$	Judgment score of item at position $j \in \tau$
g_{dis}	Disadvantaged group
g_{adv}	Advantaged group

Table 1: Notation table

bro 2020) and by designing algorithms that rely on latent feature representations instead of explicit demographic information (Hashimoto et al. 2018; Lahoti et al. 2020). Zhang and Long (2021) focus on fairness with incomplete data in classification and regression tasks. Their analysis relies on subselecting only data points (rows) where none of its values are missing. To analyze the importance of factoring missing data into the classifier models, Goel et al. (2021) studied fairness guarantees in the training procedures under various distributions. The study showed that incorporating data "missingness" can help determine the choice of fairness design principles to use in practice. To tackle information inefficiency, Noriega-Campero et al. (2019) proposed to acquire information based on the need of the group in fair classification. They showed that this helped achieve major fairness objectives, for example, equal opportunity. However, these and other algorithms as well as their empirical evaluations predominantly focus on either *fair classification* (Hashimoto et al. 2018; Celis et al. 2021; Wang et al. 2020; Mozannar, Ohannessian, and Srebro 2020; Ghosh, Kvitca, and Wilson 2023) or are restricted to re-ranking (i.e., re-ranking an existing ranking) (Ghosh, Dutt, and Wilson 2021; Mehrotra and Vishnoi 2022). While Ghosh, Dutt, and Wilson (2021) explore dealing with uncertainty in fair ranking algorithms and, in a later study, for fair classification (Ghosh, Kvitca, and Wilson 2023), the relative performance of alternate fair LTR strategies in the presence of unknown and inferred demographic groups remains an open question.

3 Experimental Methodology

We introduce fair ranking algorithms, and then describe how we compose these algorithms into a spectrum of alternate strategies for integrating fairness into fair-learning-to-rank pipelines. Next, we present tools for inferring protected attributes, followed by metrics for fairness and utility.

3.1 Preliminaries

To train an ranking model, we start with a list $C = \{c_1, ..., c_n\}$ where each candidate item c_i is associated with an attribute score vector x_i and a ground-truth relevance score $s(c_i)$ (a.k.a., judgment score). These and other useful notation is presented in Table 1.

3.2 State-Of-The-Art Fairness Interventions

In this section, we introduce the specific instantiations of the fair LTR models and the fair re-ranking models that we study, namely, DELTR (Disparate Exposure In Learning-to-Rank) and DetConstSort (Geyik, Ambler, and Kenthapadi 2019), respectively. We however, begin by introducing a fairness-unaware model, Listnet (Cao et al. 2007).

Listnet Proposed by Cao et al. (2007), it defines a loss function based on the "top one probability", defined as the probability for an item to be ranked at the top given the scores of all items. Given a list C with corresponding x_i and $s(c_i)$ values, the model is trained to assign judgment scores to unseen candidate items with attribute scores. The judgment scores can then be used to rank the items in relative order of relevance.

DELTR Proposed by Zehlike and Castillo (2020), we choose it to represent fairness-aware LTR models. This method aims to reduce disparate exposure (a measure of unfairness), while simultaneously reducing rank prediction errors. Conceptually, given a list C with corresponding x_i and $s(c_i)$ values, it is assumed that each candidate belongs to one of two disjoint groups, one of which is *protected* (g_{dis}). A group with higher visibility at the top of the ranking than another is said to have a higher exposure. This method also assumes that disparate exposure is experienced by g_{dis}. The model is trained to reduce unfairness, while aiming to maintain accurate score predictions. During training, DELTR learns to assign judgment scores to candidates using the candidates's attributes (including their protected attributes, e.g., sex) that are provided to the model. The trained model can then assign new judgment scores to unseen candidate items based on their corresponding attributes.

DetConstSort A fair re-ranking algorithm that works to improve fairness by enforcing group representation within the top k positions of a ranking. Given a list of candidates ranked by their predicted scores and a list of groups $\mathcal{G}$, DetConstSort re-ranks the list such that, for all groups $g \in \mathcal{G}$ and for all k representing a position on the ranking, the number of candidates in group g among the top k results is at least $\lfloor p_g \times k \rfloor$, where p_g is a target proportion of candidates from group g. Most commonly, the target population corresponds to the underlying distribution P_C of the candidate set. Unlike DELTR, DetConstSort is a deterministic algorithm that does not require training.

3.3 Spectrum of Fair LTR Strategies

Next, we describe the spectrum of alternate strategies we study for integrating fairness into popular LTR algorithms when the protected attribute is not known at test time.

Fairness Strategies Table 2 presents comprehensive details regarding the training, testing (ranking), and re-ranking aspects associated with each strategy. For each strategy, the model *training* may or may not include a protected attribute. For models trained with the protected attribute, we assume that the protected attributes are not available *during testing*. This leaves two possibilities for imputation during ranking and re-ranking: either (i) *inferring* the protected attribute or (ii) *hiding* it which means applying the model without gaining access to the value.

We consider three possible ways of generating the input

Fairness	Strategies	Protected Attribute Use		
		Training	Testing	Re-ranking
Unaware	**OBLIVIOUS**	n/a	n/a	n/a
	LTR	ground truth	inferred	n/a
	HIDDEN	ground truth	hidden	n/a
Aware	**FAIRLTR**	ground truth	inferred	n/a
	OBLIVIOUS+FAIRRR	n/a	n/a	inferred
	LTR+FAIRRR	ground truth	inferred	inferred (same)
	HIDDEN+FAIRRR	ground truth	hidden	inferred

Table 2: Fair ranking strategies and how they use the protected attribute **during training, testing (ranking) and re-ranking**: inferred (via noise model or inference tool), hidden (attribute replaced by constant value for all candidates), n/a.

ranking provided to the fair re-ranking algorithm, DetConstSort: (i) using a ListNet model trained without the protected attribute; or (ii) training a ListNet model using ground-truth protected attribute values and, during testing, either inferring the protected attribute or (iii) hiding it by replacing it with a constant value for all test candidates. For comparison, we also consider the three cases when the rankings were generated without the re-ranking performed by DetConstSort. These variants and DELTR altogether sum up to a total of seven strategies. We describe them in detail next; while introducing their acronyms seen in Table 2.

OBLIVIOUS This baseline stands for the approach where, *during both training and testing, a fairness-unaware ranking model has no access to the protected attributes*. The model is thus said to be "oblivious" to the protected attributes. In this setting, we use the fairness-unware Listnet.

LTR In this approach, a *fairness-unaware model is trained with access to the ground-truth protected attribute*, and *during testing it relies on inferred protected attributes*. ListNet is also used for this model. Note that the underlying ranking model Listnet does not consider fairness. Hence, we utilize this approach in isolation as another fairness-unaware baseline (different from OBLIVIOUS).

HIDDEN This strategy is to train *a fairness-unaware model with access to the ground-truth protected attributes*, and *to hide the protected attributes during testing*. For this model we also use Listnet. This approach differs from LTR because instead of inferring the protected attribute, HIDDEN neutralizes any impact of the attribute's group value on the ranking decision by replacing it by the same constant value for all candidates. This causes the model to ignore the direct effect of the protected attribute value. This approach does not rely on actual or inferred protected attribute values and is thus invariant to inference errors.

FAIRLTR In this approach, *a fairness-aware model is trained with ground-truth protected attributes* and *during testing, the protected attributes are inferred*. For this we use the DELTR model.

OBLIVIOUS+FAIRRR This stands for the approach where a *fairness-unaware ranking model trained without the protected attribute* is used to rank the list. Then *this list is processed by a fair re-ranking algorithm that uses an inferred protected attribute*. We use Listnet as the fairness-unaware ranking model and DetConstSort as the fair re-ranking algorithm. As stated in Section 3.3, DetConstSort relies on group proportions to ensure fairness, for this we use the inferred group proportions.

LTR+FAIRRR In this approach, *a fairness-unaware ranking model* (trained with ground-truth protected attributes, but using inferred attribute during testing) is *followed by a fair re-ranking algorithm which also works with the same inferred protected attributes*. We utilize ListNet for the first step, and DetConstSort as the second. As above, DetConstSort uses the inferred group proportions. This approach is similar to OBLIVIOUS+FAIRRR, however the fairness-unaware model used in the first stage is trained in the presence of the protected attributes.

HIDDEN+FAIRRR This approach utilizes *a fairness-unaware ranking model* (trained with protected attributes that are then hideen during testing) followed by a *fair re-ranking algorithm which also works with the same inferred protected attributes*. We also utilize ListNet for the first step, and DetConstSort as the seond step. As in OBLIVIOUS+FAIRRR and LTR+FAIRRR, DetConstSort uses the inferred group proportions.

3.4 Inference for Missing Protected Attributes

Inference services use other available attributes of the candidates, such as names, email, or images, to infer demographic attributes such as sex, age, or race of a person. While many such services exist, we characterize three popular solutions for inferring sex used in our experiments. In Section 4.5, we evaluate their accuracy on the datasets studied in this paper.

Behind The Name This service[1] gives the user the option to use only the first name or the full name to deduce sex. It utilizes the etymology (meaning) of a name and history of names to infer sex. It covers various regions of the world with names collated from national statistics agencies.

Gender API This service[2] uses either the first name or a combination of both first and last names to deduce sex. The

[1] https://www.behindthename.com/

[2] https://gender-api.com/

system can also leverage supplementary parameters, such as an email address or location (country, IP address, and browser) along with publicly accessible governmental data, and social media data.

Namsor This service[3] uses both the first and last name to infer sex. It claims a broad coverage across all languages, alphabets, countries, and regions. The foundational data is sourced from a compilation of 1.3 million names extracted from baby name statistics, encompassing various countries, morphology, languages, and ethnicity.

Names not Recognized By Inference Service. Each inference service returns candidates whose protected attribute could not be determined with a degree of certainty above a threshold, here called unknowns. Since these attributes are required by the fairness-aware strategies, we will need to assign some valid value to these *unknowns* (see Section 4.5).

3.5 Metrics of Fairness and Utility

We employ the following metrics in our analysis, encompassing both established fairness and utility measures.

Rank Fairness Metrics We select three primary metrics to assess rank fairness, two from a representation (Skew and NDKL) and one from an exposure standpoint.

- *Skew* (Geyik, Ambler, and Kenthapadi 2019). The skew is a fairness metric used for determining the (dis)advantaged group. We assume that candidates benefit more from being at the top of the ranking. Skew is defined for a group $g \in \mathcal{G}$ in ranking τ measured at position k as:

$$Skew_g(\tau)@k = \frac{p_{\tau@k,g}}{p_{C,g}}, \tag{1}$$

  where $p_{\tau@k,g}$ is the proportion of items belonging to g within the top k items in ranking τ, and $p_{C,g}$ the proportion of items belonging to g in the entire ranked candidate set C. A skew value of 1 is best and indicates that the group g's proportion of the top k positions in τ is the same as its proportion in the item set C. Values above 1 indicate group g is over-represented and values below 1 indicate group g is under-represented.
- *NDKL* (Geyik, Ambler, and Kenthapadi 2019). The Normalized Discounted Kullback-Leibler divergence of a ranking τ given a set of groups $\mathcal{G}$ is:

$$NDKL@k(\tau) = \frac{1}{Z}\sum_{i=1}^{k}\frac{1}{\log_2(i+1)}d_{KL}(P_{\tau@k}||P_C), \tag{2}$$

  where $d_{KL}(P_{\tau@k}||P_C)$ is the KL-divergence between $P_{\tau@k}$, the (discrete) distribution of group proportions in the first k positions of τ and P_C, the distribution of group proportions in the entire item set C. Then $Z = \sum_{i=1}^{k}\frac{1}{\log_2(i+1)}$. NDKL ranges from 0 to ∞, where lower value of zero indicates that, at all prefixes of ranking τ, all groups are represented proportionally. NDKL assesses fairness across all groups and does not indicate which group is over- or under-advantaged.

[3]https://namsor.app/

- *Average Exposure* (Singh and Joachims 2018) and Exposure Ratio (Zehlike and Castillo 2020). The DAdv/Adv Exposure Ratio of a group g_{dis} relative to another group g_{adv} in ranking τ is:

$$DAdv/Adv\ Exp\ Ratio(\tau) = \frac{AvgExposure(\tau, g_{dis})}{AvgExposure(\tau, g_{adv})}, \tag{3}$$

  where Average Exposure for group g in ranking τ is $AvgExposure(\tau, g) = \sum_{c_i \in g} Exposure(\tau, c_i)/|g|$ and the exposure of item c_i in ranking τ is $Exposure(\tau, c_i) = 1/\log_2(\tau(c_i)+1))$.
  The ideal DAdv/Adv Exp-Ratio is 1 indicating both groups have the same average exposure in ranking τ. A ratio below 1 means group g_{dis} is under-exposed in τ (i.e., unfairly disadvantaged) and a value above 1 means g_{dis} is over-exposed (i.e., unfairly advantaged) in τ.

Utility Metric. The utility of a ranking captures the relevance of the ordered items with respect to a criterion.

- *NDCG* (Järvelin and Kekäläinen 2002). The Normalized Discounted Cumulative Gain of a ranking τ is given as:

$$NDCG@k(\tau) = \frac{1}{Z}\sum_{i=1}^{k}\frac{s_\tau(i)}{\log_2(i+1)}, \tag{4}$$

  where $s_\tau(i)$ is the score of the i-th element in the ranked list τ and $Z = \sum_{i=1}^{k}\frac{1}{\log_2(i+1)}$. NDCG gives a sense of the order the documents in terms of their relevance. A value of 1 denotes the ranking has the highest utility (i.e., it orders items by decreasing scores) and as NDCG decreases down to 0, τ provides less utility.

4 Experimental Design

4.1 Data Sets

We train our models using real-world datasets containing the ground truth for the protected attribute "sex". Before training, we randomly split our data into training and testing (ranking) datasets using an 80/20 split, respectively. In the controlled study described in Section 4.4, four real-world datasets – Law, NBA/WNBA, COMPAS and Boston Marathon – are used in the experiments. Since the Law dataset lacked attributes (names) suitable for inferring the protected attribute, we have only used the other three data sets in the second study described in Section 4.5.

Law This dataset was obtained from the original DELTR (Zehlike and Castillo 2020) experimental repository. It was initially derived from a study conducted by Wightman (1998) to assess potential bias against minorities in LSAT scores. It consists of anonymized information from first-year students at some law schools.

COMPAS This dataset was collected by Propublica in their analysis of COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) tool (Larson et al. 2016). It contains attributes used for predicting the likelihood for a criminal defendant to offend again.

(W)NBA This dataset contains information about WNBA (Women's National Basketball Association) and NBA (National Basketball Association) players. Each player is associated with career points which determine their position in the ranking. Other attributes include number of seasons they played in the career and their average player efficiency ratio.

Boston Marathon This dataset contains a list of Boston Marathon finishers in 2019. It contains both male and female runners. Each runner's times at 7 different stages were used in our experiment [4].

Table 3 shows the number of males and females in each ranking dataset during deployment (testing). Each test set represents 20% of the respective complete dataset.

4.2 Disadvantaged Groups in Our Datasets

To identify the disadvantaged group within each of our datasets when candidates are ordered by ground-truth judgment scores, we compute the *skews* for each group at every position in the ranking (Equation 1). Graphs displaying the distribution of these groups throughout each of the datasets are included in the supplemental material of the extended version (Olulana et al. 2024, Fig. 1). The group displaying lower representation at the top of a ranking is categorized as the disadvantaged group.

4.3 Model Training and Parameter Settings

For each dataset listed in Section 4.1, we consider fairness-unaware models trained without the protected attributes, fairness-unaware models trained reliant on the protected attributes, and fairness-aware models requiring the protected attributes. Additional information regarding the training of models has been provided in Section 4.3 from the supplemental material.

4.4 Controlled Study: Varying Inference Errors

In this first study, we systematically vary the level of errors in the protected attributes in our ranking scenarios, mimicking errors commonly encountered in demographic inference. Specifically, we consider three cases: bidirectional errors ($g_{dis} \leftrightarrow g_{adv}$), i.e., some disadvantaged candidates are incorrectly inferred as advantaged and vice-versa, and unidirectional errors, where only candidates of one of groups are mistaken by candidates of the other ($g_{dis} \rightarrow g_{adv}$ and $g_{dis} \leftarrow g_{adv}$). Unidirectional errors capture the situation where candidates whose group could not be inferred are all assigned to either g_{adv} (or g_{dis}). We control the error level using a parameter ϵ defined as the fraction of candidates in each group for which inference is wrong. The results for the unidirectional scenarios are discussed in the supplemental material.

For each of these scenarios, we generate controlled test sets by varying ϵ from 0% to 100% in increments of 10%. To reduce the variance in the results, we increase the error levels cumulatively, i.e., all candidates whose protected attribute was flipped at one error level will retain the wrong value at a higher error level. This process is repeated 5 times with different random seeds for ϵ from 10% to 90%. Noting that for ϵ equal to 0% and 100% the results remain the same, this yields 47 test scenarios for each dataset. We employ these scenarios as a test benchmark for the approaches previously described in Section 3.3. We compute fairness metrics using ground-truth labels. We average over all ϵ values. Note that $\epsilon = 0\%$ serves as a reference point, as it corresponds to using the ground-truth protected attribute values.

[4] kaggle.com/datasets/daniboy370/boston-marathon-2019

4.5 Case Studies: Using Popular Inference Methods on Real-World Data Sets

In this second study, we use the real-world inference services described in Section 3.4 to obtain demographic attributes for the three real-world datasets introduced in Section 4.1.

Table 4 shows the accuracy of each inference service when provided with candidates' names from our datasets. Each inference service has a subset of candidate items for which it cannot identify a sex. In such cases, the inference service returns an "unknown" value. Gender API has the lowest number of unknown values, whereas Namsor has the highest. When considering the accuracy for only the identified items, Namsor performs best.

Handling Unrecognized Names Each inference service returned some "unknowns", i.e., candidates' sex could not be inferred. Each of these candidates was assigned to the disadvantaged group g_{dis}. This design choice would prevent a disadvantaged candidate to be misclassified as part of g_{adv} and thus receive a penalty when fairness-aware strategies are in place. Table 4 shows the number of candidates without inference results. It is important to note that the final error rate changes due to adding the unrecognized names to a group.

4.6 Measuring Fairness and Utility

To calculate the fairness metrics, we use the ground-truth sex attribute to capture true fairness levels. For utility, we use the ground-truth judgment scores. To analyze the effect of wrong demographic inference on the fairness and utility measures of rankings generated by fairness-aware algorithms, we follow the steps sketched below.

Step 1. We measure the skews (see Definition in Section 4.2) of each group on the ranking directly obtained from judgment scores, i.e., before applying any ranking algorithm. For each dataset, the group that has the lowest skew for all or nearly all positions at the top of the ranking is defined as the disadvantaged group g_{dis} (see Table 3).

Step 2. We train each ranking model as in Section 4.3.

Step 3. Lastly, we apply each model to a test set (according to each strategy in Section 3.3), and compute fairness Dadv/Adv exposure ratio and NDKL and utility NDCG metrics.

In the controlled error rate experiments, we average the metrics over the five variants of the test set generated for each ϵ.

5 Results and Analysis

We present our results for the 1^{st} simulation scenario of the controlled inference experiments in Section 5.1 and our use cases with popular inference services on real-world data sets

Dataset	Training Size	Test Males	Test Females	Test Size	g_{dis}	Inference attribute	Study used in Controlled	Study used in Case	Target feature
Law	4,882	55%	45%	1,221	females	not available	✓	✗	first year scores
(W)NBA	3,726	78%	22%	992	females	names	✓	✓	career points
COMPAS	4,917	82%	18%	1,257	males	names	✓	✓	recidivism score
Boston M	21,063	55%	45%	5,226	females	names	✓	✓	official time

Table 3: Statistics of the training and test data, disavantaged group designation, attribute used for inference and set of experiments in which each data set was used.

	Service	Test Size	Have Inference Result Correct	Incorrect	Total	Inference Result Not Available g_{dis}	g_{adv}	Total
(W)NBA	**Gender API**	992	931 (94%)	39 (4%)	970 (98%)	7 (.7%)	15 (2%)	22 (2%)
	Behind The Name		746 (75%)	8 (1%)	754 (76%)	81 (8%)	157 (16%)	238 (24%)
	Namsor		397 (40%)	0 (0%)	397 (40%)	131 (13%)	464 (47%)	595 (60%)
COMPAS	**Gender API**	1,257	1135 (90%)	72 (6%)	1207 (96%)	39 (3%)	11 (1%)	50 (4%)
	Behind The Name		905 (72%)	38 (3%)	943 (75%)	239 (19%)	75 (6%)	314 (25%)
	Namsor		500 (40%)	15 (1%)	515 (41%)	608 (48%)	134 (11%)	742 (59%)
BostonM	**Gender API**	5,226	5019 (96%)	155 (3%)	5174 (99%)	23 (.4%)	29 (1%)	52 (1%)
	Behind The Name		4302 (82%)	88 (2%)	4390 (84%)	443 (8%)	394 (8%)	836 (16%)
	Namsor		2090 (40%)	0 (0%)	2090 (40%)	1474 (28%)	1662 (32%)	3136 (60%)

Table 4: Statistics of inferred protected attribute for each data set and service. Candidate items with inferred protected attributes are grouped by inference correctness. Candidate items for failed inference are grouped based on ground-truth attribute values.

in Section 5.2. Results for the 2$^{\text{nd}}$ and 3$^{\text{rd}}$ simulation scenarios are available in the supplemental material of extended version (Olulana et al. 2024, Figs. S7, S8).

5.1 Controlled Inference Error Evaluation

Figure 1 shows our results in terms of the fairness and utility measures for the datasets when we have applied controlled inferencing processes to infer protected attributes. For fairness, we present Dadv/Adv exposure ratio plots and NDKL plots. For utility, we present NDCG@100.

OBLIVIOUS OBLIVIOUS is represented in Figure 1 by a horizontal line, as it is invariant to the inference error ϵ.

- OBLIVIOUS consistently has Dadv/Adv exposure ratio farther from 1.0 than the fair interventions for up to 15-20% error across all the datasets. This is expected, as the model does not try explicitly optimize fairness. This shows that if the error is small enough, fairness interventions are a sufficient and safe option for practitioners (as opposed to using an oblivious model). Moreover, it consistently yields similar Dadv/Adv exposure ratio values to HIDDEN, showing that hiding the protected attributes during testing would yield fairness levels to those when the protected attribute is not used during training.
- Opting for OBLIVIOUS when subject to possible errors in the protected attribute values also proves to be a better option than choosing LTR in terms of Dadv/Adv exposure ratio for up to 15 - 20% error rate. This is expected, as including the protected attributes in an already biased ranking only reinforces the bias. This implies that, disregarding the protected attribute during training can improve fairness even though it is not explicitly optimize.
- For NDKL, OBLIVIOUS's value is similar to those of the reranking fairness interventions' across different error rates except for in the (W)NBA dataset, where OBLIVIOUS does better.
- For NDCG, in cases where the Dadv/Adv exposure ratio were lower than those of the fairness strategies, OBLIVIOUS had higher NDCG values and vice versa (except for the LAW dataset). This is as expected due to the fairness-utility trade-off.

LTR

- For LTR, the Dadv/Adv exposure ratio rises with the inference error across all datasets until it eventually converges. This is in line with our expectations, as errors in demographic attributes lead the model to misidentify the disadvantaged individuals as advantaged, resulting in a heightened degree of preference towards the disadvantaged group, g_{dis}.
- Interestingly, while fairness is not an inherent objective of the model, group fairness emerges as a byproduct of erroneous inferences throughout the rankings. However, it is worth noting that for $\epsilon \geq 30\%$, the initial disparity is reversed and the advantaged group gets less exposure. The only exception is (W)NBA, where the effect of the sex coefficient is relatively small (as seen by the small difference in Dadv/Adv exposure ratio between LTR at $\epsilon = 0\%$ and HIDDEN).
- NDKL decreases (increased fairness) with inference error, however, there is a turning point after which the orig-

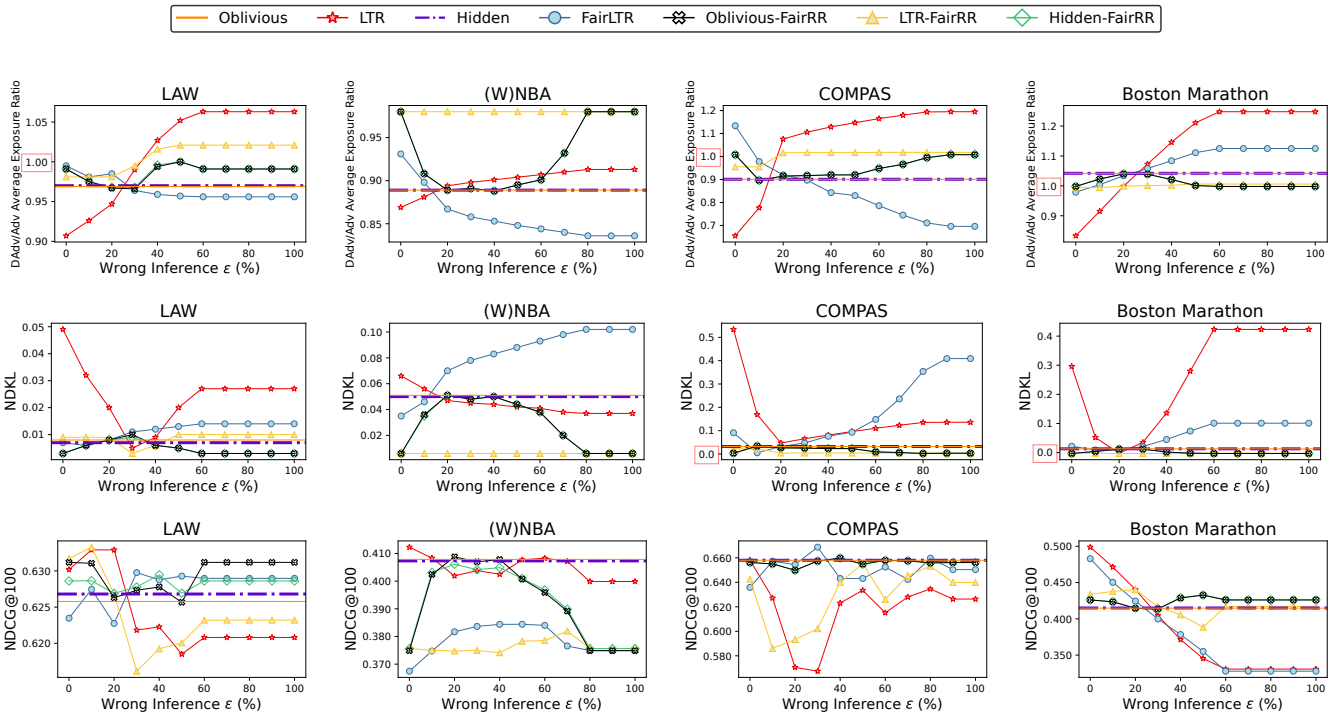

Figure 1: DAdv/Adv Exposure Ratio, NDKL & NDCG@100 graphs for the 1st simulation scenario ($g_{dis} \leftrightarrow g_{adv}$). Ideal values are highlighted with red boxes on the y-axes whenever visible.

inal advantaged group becomes severely disadvantaged. This is in line with results for Dadv/Adv exposure ratio.

- NDCG tends to decrease with inference noise, as advantaged (resp. disadvantaged) candidates move down (resp. up) the ranking. This effect is more pronounced in the COMPAS and Boston Marathon data sets because the coefficient associated with the protected attribute has a large magnitude.

HIDDEN Represented in Figure 1 by a horizontal dashed line, it is invariant to the inference error ϵ.

- In the cases where the inference error rate is up to 15-20%, this method is outperformed by all the fairness strategies, which yield Dadv/Adv exposure ratio closer to 1.0. In contrast, for higher error rates, the effectiveness of the fairness strategies drops significantly.
- We observe that the NDKL is consistently outperformed by all the fairness strategies for up to 15-20% inference error rates, which is similar to the Dadv/Adv exposure ratio staying close to 1.0 in our results.
- The NDCG of HIDDEN tends to be higher (i.e., better) than those of the fair strategies, because it does not attempt to compromise utility and fairness.

FAIRLTR

- For the fair LTR fairness strategy, FAIRLTR, Dadv/Adv exposure ratio decreases (decreased fairness) with the error rate across the first three datasets in all scenarios. When candidates in g_{dis} are incorrectly identified as part of g_{adv}, they do not get the score boost the model would have given them otherwise, therefore decreasing their exposure. The only exception, Boston Marathon, is easily explained by inspecting the model coefficients: although smaller in magnitude than in LTR, FAIRLTR still yields a negative coefficient associated with being part of the disadvantaged group. Hence, as the inference errors increase, disadvantaged candidates benefit.
- NDKL tends to always increase (decreased fairness) and then converge. This is because NDKL summarizes unfairness across groups – i.e., its value does not reveal which group is disadvantaged. Whereas Dadv/Adv exposure ratio quantifies which groups is over- or under-advantaged.
- NDCG tends to increase as advantaged (resp. disadvantaged) candidates get a bonus (resp. penalty) in score due to being incorrectly inferred as part of the other group.

OBLIVIOUS+FAIRRR

- For OBLIVIOUS+FAIRRR, there is not much variation in the Dadv/Adv exposure ratio on the Law and Boston Marathon datasets. The value initially reduces in the fairness value (away from 1.0), but eventually returns to its inital value. This is due to the fact that, from the DetConstSort's perspective, the group proportions between g_{dis} and g_{adv} have been swapped. For instance, in the (W)NBA case, at $\epsilon = 100\%$, DetConstSort perceives

78% of the population as female (i.e., the inverse of Table 3). This will ensure that this proportion is present at any top k positions in the ranking. However, since the sex attribute values are completely swapped, but candidates' scores remain the same, DetConstSort will put the same candidates in the same positions as had the inferred attributes been 100% accurate. Thus it recovers the exact same ranking as in $\epsilon = 0$.

- Consequently, NDKL does not change much with error for OBLIVIOUS+FAIRRR. Even in (W)NBA, the NDKL increase is relatively small: its maximum is 0.05.
- NDCG does not vary much for LAW, COMPAS and Boston Marathon. The NDKL returns to the same value as for $\epsilon = 0\%$ as the error approaches 100%, since the two rankings are identical.

LTR+FAIRRR

- Surprisingly, for LTR+FAIRRR, Dadv/Adv exposure ratio does not exhibit much variation with inference error in any dataset. After careful investigation, we found that this is because the increase/decrease in score caused by providing the wrong protected attribute to Listnet is often counteracted by DetConstSort when the perceived group proportions change, as we explain next. For instance, a candidate c from g_{dis} perceived by Listnet as part of g_{adv} gets a bonus in their score. Yet, these inference errors also lead to a small increase in the proportion of the advantaged group.
- The NDCG decreases and then converges across the increases error level for the LAW and Boston Marathon datasets, it is for stable for the (W)NBA dataset and less stable for COMPAS.

HIDDEN+FAIRRR

The results are identical to OBLIVIOUS+FAIRRR for the Dadv/Adv exposure ratio and NDKL. This is expected, following the similarity between OBLIVIOUS and HIDDEN. Slight differences are however observed in the NDCG values for the LAW and (W)NBA datasets.

5.2 Real World Data Set Use Cases with SOTA Inference Techniques

As explained in Section 3.4, we assign the candidates whose protected attribute value were unknown to the disadvantaged group. For the ease of exposition, we sort services *in increasing order of error rate*: Gender API (GAPI), Behind The Name (BTN) and Namsor (NMSOR). We add a case for ground-truth protected demographics (0% error rate), referred to as G-TRUTH in Figure 2. In general, we observe similar behavior as in the error simulation studies.

OBLIVIOUS It yields Dadv/Adv exposure ratio values farther away from 1.0 than all the fair interventions in the G-TRUTH results. Yet, it is closer to 1.0 than LTR.

LTR We observe that the Dadv/Adv exposure ratio values slowly increases from the least to the most inaccurate service, exhibiting a slight variation with smaller errors (6% for GAPI for (W)NBA) and a larger variation with larger errors (47% for NMSOR for (W)NBA). This is consistent with our results in Figure 1). NDKL and NDCG results are also in line with our controlled experiments.

HIDDEN In terms of Dadv/Adv exposure ratio, fairness strategies yield values close to 1.0 on Boston Marathon, whereas HIDDEN ends up overexposing the disadvantaged group. This is considered unfair as it deviates from the ideal Dadv/Adv exposure ratio value. Therefore, HIDDEN is outperformed by the fairness strategies in terms of fairness. Conversely, regarding NDCG, HIDDEN matches or exceeds the performance of fairness strategies on (W)NBA and COMPAS. Due to the special nature of the Boston Marathon dataset (where the disadvantaged group is overexposed by HIDDEN), HIDDEN leads to the lowest utility among all strategies.

FAIRLTR Fairness decreases (Dadv/Adv exposure ratio deviates further from 1.0) observe with an increase in error rates in the (W)NBA dataset and in Boston Marathon. COMPAS is an exceptional case since FAIRLTR overexposes the disadvantaged group, so an increase in error brings Dadv/Adv exposure ratio closer to 1.0. NDKL mirrors the same fairness trends as in the Dadv/Adv exposure ratio values. For COMPAS and Boston Marathon, NDCG are the smallest values for the most inaccurate service (NMSOR).

OBLIVIOUS+FAIRRR The larger the error in inference, the father away from 1.0 the Dadv/Adv exposure ratio value is. For NDKL, the deviation from the perfect representation as the error increases is not clearly seen except in (W)NBA dataset. NDCG values reflect the fairness-utility trade-off; higher NDCG values for lower Dadv/Adv exposure ratio values and vice-versa.

LTR+FAIRRR Dadv/Adv exposure ratio again stays relatively constant across the inference services. In terms of NDCG, this strategy achieves consistent performance across different services.

HIDDEN+FAIRRR Dadv/Adv exposure ratio stays relatively constant with increases in inference error rate, corroborating the results from the controlled experiments. Observed trends for NDCG are also as expected: they are somewhat invariant to inference errors, exhibiting only a minor increase with error in (W)NBA dataset.

6 Discussion: Insights and Take-Aways

Our paper focuses on how errors in inferring the protected attribute may affect the fairness and utility metrics of rankings produced by a wide spectrum of alternate LTR-based ranking systems. These methods can instill fairness at one of two possible stages: as part of a Fair LTR model or a posteriori by using a Fair ReRanker. We investigated the feasibility and implications of inferring or neglecting protected attributes when they are not available. We conclude:

- In scenarios characterized by high inference error rates, fairness models that require protected attributes, while still effective in increasing the disadvantaged group exposure, may inadvertently lead to the advantaged group having lower exposure than the former

  (e.g, Boston Marathon in Fig. 1).

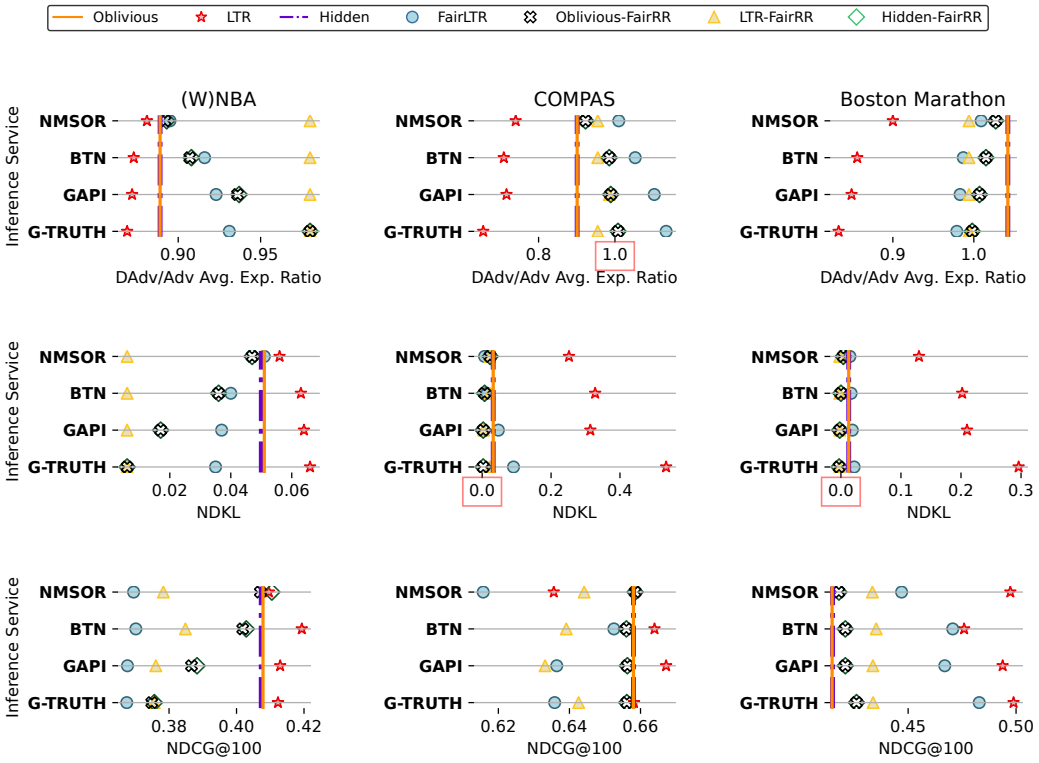

Figure 2: DAdv/Adv Exposure Ratio, NDKL and NDCG@100 graphs for all inference services and LTR strategies.

- The impact of inference errors on fairness varies depending on the model employed. Notably, fair LTR models such as DELTR exhibit distinct behavior compared to fairness-unaware models like ListNet followed by a Fair ReRanker. The latter demonstrates superior capability in maintaining fairness when faced with higher levels of inference errors. While LTR boosts the scores of candidates incorrectly inferred as advantaged, DetConstSort prevents overexposure of this group by enforcing that both candidates inferred as advantaged or disadvantaged are proportionally represented at all ranking cutoffs.
- We studied whether to hide or infer protected features if they are not readily available. For the LTR strategy, we observated that wrong inference can improve fairness, as candidate items from the disadvantaged group typically get a better rank when misclassified. Nonetheless, the takeaway is that inference methods are often unpredictable, and their use should be heavily monitored, if not discouraged. Both our experiments with controlled inference error rates and real world use cases with state-of-the-art inference techniques corroborate that finding.
- FAIRLTR strategies like DELTR maintain fairness even with imperfect inference, with better fairness achieved as inference accuracy increases. Conversely, strategies that do not depend on inferring protected attributes, like HIDDEN may provide lower fairness metrics overall.

Our research emphasizes the dangers of demographic inference for practitioners along with other important insights.

Study Limitations and Future Work. While our study focuses on gender, other demographics like race or religion along with their inferencing services also are needed to reach a general recommendation for practitioners. Demographic attributes can be multi-valued, while our investigation, similar to prior work (Ghosh, Dutt, and Wilson 2021), focused the core binary scenario.

Conclusion

The absence of demographic features may affect the effective functioning of fair ranking systems. Thus, practitioners may attempt to overcome this by either promoting fairness despite their absence or turning to demographic inference tools to attempt to infer them. Our study sheds some light on this decision, namely, we find that if this inference is deemed necessary and inevitable, fair re-ranking solutions serve as a more resilient alternative compared to Fair LTR solutions.

Acknowledgments

This work is supported in part by NSF grants IIS 2007932, CNS-1852498, IIS-1910880, CSSI-2103832, and AAUW. Thank you to undergraduate REU/MQP researchers Pietrick, Romportl, Smith, Tessier, Vadlamudi, and Venkataraman for contributions to early stages of this project. Thanks Daisy research group for valuable feedback.

Ethical Statement

Ethical Considerations We worked with datasets, AI models, and inference tools that are all publicly available. As such, we adhered to all licensing agreements and download requirements related to their use. Our experimental approach is explained in Section 3.

For the datasets, we do not display any information about these datasets at the granularity of individuals in our work - only at the aggregated level of demographic groups. Licensing details about the models and inferencing tools are found in the following: DELTR (github.com/fair-search/fairsearch-deltr-python/blob/master/LICENSE), GenderAPI (gender-api.com/en/terms-of-use), Behind The Name (behindthename.com/info) and Namesor (namsor.app/terms-and-conditions).

All source code we produced and experimental artifacts of our experimental study will be made available on *github*, upon acceptance of this paper.

Adverse Impacts While our research finds that using inaccurate demographic inference tools is typically inadvisable, we are mindful that fairness is contextual to the problem, tools, and task at hand. Practitioners should continuously monitor the fairness performance of ranking models and as well as assess whether using inference methods in place of demographic data has unintended consequences.

Understanding the ramifications of using demographic inference tools, as done in our work, should in no way be construed as endorsing such practices. Rather, with this work, we aim to uncover the potential drawbacks and benefits that these practices may cause so that practitioners and policy makers can make informed decisions.

Researcher Positionality As computer scientists and data scientists by training, we are influenced by our perspectives on how we approach research problems and how we develop and/or study computational solutions to fairness problems in our society related to the introduction of digital tools.

References

BehindTheName. 1996. https://www.behindthename.com/. Last Accessed: January 21, 2024.

Bogen, M.; Rieke, A.; and Ahmed, S. 2020. Awareness in practice: tensions in access to sensitive attribute data for antidiscrimination. In *ACM FAT**, 492–500.

Cao, Z.; Qin, T.; Liu, T.-Y.; Tsai, M.-F.; and Li, H. 2007. Learning to rank: from pairwise approach to listwise approach. In *ICML*, 129–136.

Celis, L. E.; Huang, L.; Keswani, V.; and Vishnoi, N. K. 2021. Fair classification with noisy protected attributes: A framework with provable guarantees. In *ICML*, 1349–1361.

Cheng, N.; Chen, X.; Chandramouli, R.; and Subbalakshmi, K. 2009. Gender identification from e-mails. In *IEEE SSCI*, 154–158.

Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *ITCS*, 214–226.

eCFR. 2024. Electronic Code of Federal Regulations. https://www.ecfr.gov/current/title-12/chapter-X/part-1002/subpart-A/section-1002.5#p-1002.5(b). May 9, 2024.

Ekstrand, M. D.; Das, A.; Burke, R.; and Diaz, F. 2021. Fairness and discrimination in information access systems. *arXiv preprint arXiv:2105.05779*.

Fink, C.; Kopecky, J.; and Morawski, M. 2012. Inferring gender from the content of tweets: A region specific example. In *AAAI ICWSM*, 459–462.

Gender-API. n.d. Gender-API (Version 2). https://gender-api.com. Last Accessed: January 21, 2024.

Geyik, S. C.; Ambler, S.; and Kenthapadi, K. 2019. Fairness-aware ranking in search & recommendation systems with application to LinkedIn talent search. In *ACM SIGKDD*, 2221–2231.

Ghazimatin, A.; Kleindessner, M.; Russell, C.; Abedjan, Z.; and Golebiowski, J. 2022. Measuring fairness of rankings under noisy sensitive information. In *ACM FAccT*, 2263–2279.

Ghosh, A.; Dutt, R.; and Wilson, C. 2021. When fair ranking meets uncertain inference. In *ACM SIGIR 2021*, 1033–1043.

Ghosh, A.; Kvitca, P.; and Wilson, C. 2023. When Fair Classification Meets Noisy Protected Attributes. In *AAAI/ACM AIES*, 679–690.

Goel, N.; Amayuelas, A.; Deshpande, A.; and Sharma, A. 2021. The importance of modeling data missingness in algorithmic fairness: A causal perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 7564–7573.

Hashimoto, T.; Srivastava, M.; Namkoong, H.; and Liang, P. 2018. Fairness Without Demographics in Repeated Loss Minimization. In *ICML*, 1929–1938.

Holstein, K.; Vaughan, J. W.; Daumé, H.; Dudík, M.; and Wallach, H. M. 2018. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? In *ACM CHI*.

Järvelin, K.; and Kekäläinen, J. 2002. Cumulated gain-based evaluation of IR techniques. *ACM TOIS*, 20(4): 422–446.

Jillson, E. 2021. Aiming for truth, fairness, and equity in your company's use of AI. *Federal Trade Commission*.

Kırnap, Ö.; Diaz, F.; Biega, A.; Ekstrand, M.; Carterette, B.; and Yilmaz, E. 2021. Estimation of fair ranking metrics with incomplete judgments. In *The Web Conference*, 1065–1075.

Lahoti, P.; Beutel, A.; Chen, J.; Lee, K.; Prost, F.; Thain, N.; Wang, X.; and Chi, E. H. 2020. Fairness without Demographics through Adversarially Reweighted Learning. *ArXiv*, abs/2006.13114.

Larson, J.; Mattu, S.; Kirchner, L.; and Angwin, J. 2016. How we analyzed the COMPAS Recidivism Algorithm.

Li, Y.; Chen, H.; Fu, Z.; Ge, Y.; and Zhang, Y. 2021. User-oriented fairness in recommendation. In *Proceedings of the Web Conference 2021*, 624–632.

Li, Y.; Chen, H.; Xu, S.; Ge, Y.; Tan, J.; Liu, S.; and Zhang, Y. 2022. Fairness in recommendation: A survey. *arXiv preprint arXiv:2205.13619*.

Mehrotra, A.; and Vishnoi, N. 2022. Fair ranking with noisy protected attributes. In *NeurIPS*, 31711–31725.

Mozannar, H.; Ohannessian, M.; and Srebro, N. 2020. Fair learning with private demographic data. In *ICML*, 7066–7075.

Namesor. n.d. https://namsor.app/. Last Accessed: January 21, 2024.

Noriega-Campero, A.; Bakker, M. A.; Garcia-Bulle, B.; and Pentland, A. 2019. Active fairness in algorithmic decision making. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 77–83.

Olulana, O.; Cachel, K.; Murai, F.; and Rundensteiner, E. 2024. Hidden or Inferred: Fair Learning-To-Rank with Unknown Demographics. *arXiv preprint arXiv:2407.17459*.

Patro, G. K.; Porcaro, L.; Mitchell, L.; Zhang, Q.; Zehlike, M.; and Garg, N. 2022. Fair Ranking: A Critical Review, Challenges, and Future Directions. In *ACM FAccT*, 1929–1942. ISBN 9781450393522.

Santamaría, L.; and Mihaljević, H. 2018. Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, 4: e156.

Singh, A.; and Joachims, T. 2018. Fairness of exposure in rankings. In *ACM SIGKDD*, 2219–2228.

Singh, A.; and Joachims, T. 2019. Policy learning for fairness in ranking. *Advances in neural information processing systems*, 32.

Wang, S.; Guo, W.; Narasimhan, H.; Cotter, A.; Gupta, M.; and Jordan, M. 2020. Robust optimization for fairness with noisy protected groups. In *NeurIPS*, 5190–5203.

Wang, Y.; Tao, Z.; and Fang, Y. 2022. A Meta-learning Approach to Fair Ranking. In *ACM SIGIR*, 2539–2544.

Wightman, L. F. 1998. LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series. *ERIC*.

Zehlike, M.; Bonchi, F.; Castillo, C.; Hajian, S.; Megahed, M.; and Baeza-Yates, R. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1569–1578.

Zehlike, M.; and Castillo, C. 2020. Reducing disparate exposure in ranking: A learning to rank approach. In *The Web Conference*, 2849–2855.

Zhang, Y.; and Long, Q. 2021. Assessing fairness in the presence of missing data. *Advances in neural information processing systems*, 34: 16007–16019.