What is the Chance of Being so Unfair?

Hossein Esfandiari*
Google Research
London, UK
esfandiari@google.com

Samira HosseinGhorban
Institute for Research in Fundamental Sciences
School of Computer Science
Tehran, Iran
s.hosseinghorban@ipm.ir

Reza Hosseini Dolatabadi
Sharif University of Technology
Tehran, Iran
reza.dolatabadi256@sharif.edu

Mahdi Qaempanah
Sharif University of Technology
Tehran, Iran
mahdi.ghaempanah111@student.sharif.edu

ABSTRACT

Fairness has often been seen as an ethical concern that needs to be considered at some cost on the utility. In contrast, in this work, we formulate fairness, and especially fairness in ranking, as a way to avoid unjust biases and provide a more accurate ranking that results in improvement on the actual unbiased utility. With this in mind, we design a fairness measure that, instead of blindly forcing some approximate equality constraint, checks if the outcome is plausible in a just world. Our fairness measure asks a simple and fundamental statistical question: "What is the chance of observing this outcome in an unbiased world?". If the chance is high enough, the outcome is fair. We provide a dynamic programming algorithm that, given a ranking, calculates our fairness measure. Secondly, given a sequence of potentially biased scores, along with the sensitive feature, we provide a fair ranking algorithm based on our fairness measure. Finally, we run some experiments to understand the behavior of our ranking algorithm against other fundamental algorithms.

CCS CONCEPTS

• **Applied computing** → **Law, social and behavioral sciences**;
• **Computing methodologies** → **Ranking**.

KEYWORDS

Fairness, Fairness in ranking

ACM Reference Format:
Hossein Esfandiari, Samira HosseinGhorban, Reza Hosseini Dolatabadi, and Mahdi Qaempanah. 2025. What is the Chance of Being so Unfair?. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3701716.3715507

WWW Companion '25, April 28-May 2, 2025, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1331-6/2025/04
https://doi.org/10.1145/3701716.3715507

1 INTRODUCTION

In the past decade, fairness has become a key concept in machine learning and automated decision making. Specifically, in recommendation systems and hiring platforms, fairness means that ranking mechanisms should be unbiased and not discriminate based on demographic characteristics or other protected attributes.

The group of individuals in a ranking task is called candidates who may have sensitive attributes. The algorithmic methods that address fairness differ in the representation of candidates, the type of bias, mitigation objectives, and mitigation methods such as worldviews [4]. In response to worldviews, Friedler et al. [1] highlight the need to understand the difference between our belief about fairness and the mathematical definition of fairness. They present two views that represent the two ends of the spectrum: WYSIWYG ("what you see is what you get") and WAE ("we are all equal"). WYSIWYG assumes that what we see is nearly the same as the real properties, with just a ϵ distortion. WAE assumes that biased observations cause differences in utility distributions among the candidates.

In this work, we introduce a stochastic variant of WAE, that we refer to as *Stochastic-WAE*. Based on stochastic-WAE, we provide a fairness measure that poses a fundamental statistical question: *What is the likelihood of observing this outcome in an unbiased scenario?*. If this likelihood is high enough, we consider it fair. We present Stochastic-WAE that captures randomness while keeping it independent of sensitive data. It recognizes that probability distributions for different groups, such as females and non-females, should be the same, despite potential score gaps in specific subsets.

Given a ranking, we provide a dynamic programming algorithm that answers the above question and calculates our fairness measure. This can be used on top of other ranking algorithms to measure their fairness. Next, we design a ranking algorithm that respects our fairness measure. Specifically, we design an algorithm that, given a measure δ and given a sequence of potentially biased scores, along with the sensitive feature, provides a fair ranking with maximum possible utility such that its fairness measure is at most δ.

1.1 Problem Setting

Let Ω be the set of all possible candidates that originate from a society that is originally faced with bias. For simplicity of presentation, we focus on a single binary group with a score that includes an unknown preexisting bias against women. Let

$$id(\omega) = \begin{cases} 1 & \text{if } \omega \text{ is a woman,} \\ 0 & \text{otherwise,} \end{cases}$$

where $\omega \in \Omega$. We receive the set of candidates $C = (\omega_1, \ldots, \omega_n)$ and their corresponding biased scores $(\hat{y_1}, \hat{y_2}, \ldots, \hat{y_n})$. We define the following quantity which represents the *minority proportion*

$$p = \frac{\sum_{j=1}^{n} id(\omega_j)}{n}.$$

A permutation over the candidate sample C is called a *ranking*. Let $\tau : \{1, \ldots, n\} \to C$ be the observed-score-based ranking on C, i.e., $\hat{y}_{\tau(1)} \geq \cdots \geq \hat{y}_{\tau(n)}$, in which $\tau(i)$ is the candidate at rank i. The utility of τ based on DCG approach is defined as follows [3]:

$$U(\tau) = \sum_{i=1}^{n} \frac{\hat{y}_{\tau(i)}}{\log_2(i+1)}. \tag{1}$$

In the context of ranking algorithms, the worldview of "We are all equal (WAE)" implies that individuals with similar qualities should have an equal chance of being ranked similarly [3].

In this work, we adopt a statistical approach to similarity, and hence, we assume that the unbiased scores for men and women, in which discrimination based on gender is absent, are taken from the same unknown probability distribution. Let y_i be the i-th candidate's unbiased score which is taken from an unknown probability distribution $\mathcal{P}_Y$ where a random variable $Y : \Omega \to \mathbb{R}$ represents the unbiased score for a given candidate. Based on the statistical WAE worldview, the value of y_i is independent of the group to which the candidate ω_i belongs. Therefore, $(y_1, \ldots, y_n)$ is an independent and identically distributed (i.i.d.) random sequence. Let the permutation σ be an ordering of $y_1, \ldots, y_n$ s.t. $y_{\sigma(1)} \geq \ldots \geq y_{\sigma(n)}$. The permutation σ is an *unbiased ranking* for C.

Since the unbiased scores are taken the same distribution, we have $\mathbb{E}[Y(\omega) \mid id(\omega) = 1] = \mathbb{E}[Y(\omega) \mid id(\omega) = 0]$, and consequently, intuitively, the arrangement of women and men in cut-off points of the ranking should not be statistically rare. Next we formalize this notion.

1.2 Defining Fairness in Ranking via Statistical-WAE worldview

For a ranking $\tau : \{1, \ldots, n\} \to C$, let $\mathcal{F}_\tau = \{S_1^\tau, \ldots, S_n^\tau\}$ where $S_i^\tau = \{\tau(1), \ldots, \tau(i)\}$. We call S_i^τ the *i'th partial set corresponding to* τ. Hence, $S_1^\tau \subset S_2^\tau \subset \cdots \subset S_n^\tau$. Let X_i^τ be the number of women in S_i^τ. Sort $Y(\omega_1), Y(\omega_2), \ldots, Y(\omega_n)$ to obtain an unbiased ranking σ such that $Y(\omega_{\sigma(1)}) \geq Y(\omega_{\sigma(2)}) \geq \cdots \geq Y(\omega_{\sigma(n)})$.

Similarly, X_i^σ is the number of women in the i'th partial set S_i^σ. Since Y is a random variable, one can see that σ is a random permutation and so X_i^σ is a random variable. For the sake of simplicity, we will denote X_i^σ by X_i in the rest.

Definition 1. *For a ranking $\tau : \{1, \ldots, n\} \to C$, the partial set S_i^τ is said to be "δ-rare" if and only if the following inequality holds:*

$$Pr[X_i \leq X_i^\tau] < \delta.$$

Moreover, we say X_i^τ is in the δ-tail of $\mathcal{P}_{X_i}$.

Now, we want to formalize the notion of fairness concerning the statistical WAE worldview precisely.

Definition 2. *A ranking $\tau : \{1, \ldots, n\} \to C$ is called δ-fair if and only if none of the members of the $\mathcal{F}_\tau$ are δ-rare.*

This definition explicitly says that none of the partial sets associated with a fair ranking is in the δ-tail of $\mathcal{P}_{X_i}$. In other words, a ranking is δ-fair, if the occurrence probability of the least probable partial set is not lower than δ.

2 RANKING ALGORITHMS

In this section, we provide an algorithm to measure the fairness in ranking based on our fairness criteria. Next, we design an algorithm that fairly ranks a given candidate set.

2.1 An Algorithm to Measure Ranking Fairness

In algorithm 1, as we assumed the unbiased scores of all candidates come from the distribution $\mathcal{P}_Y$, at each step of a ranking (consider the process as a step-by-step procedure that puts the candidates in their place respectively from the first to the nth place), the probability of the next candidate to be a woman is the proportion of unranked women to the total number of remaining candidates. We prove the following theorem.

Theorem 1. *For a given candidate sample $C = (\omega_1, \cdots, \omega_n)$, let n_1 and n_2 be the total number of women and men respectively. Let a tuple (i, m) represent the event of seeing m men in the first i candidates of a fair ranking. Then, by statistical WAE worldview, the following equation holds*

$$\begin{aligned} Pr[(i,m)] = &\left(\frac{n_2 - (m-1)}{n - (i-1)}\right) Pr[(i-1, m-1)] \\ &+ \left(\frac{n_1 - (i-1-m)}{n - (i-1)}\right) Pr[(i-1, m)]. \end{aligned}$$

This theorem allows us to calculate the probability of a partial set using the probabilities of the previous step's partial sets. This enables us to develop a dynamic programming algorithm to calculate the partial set probabilities which is represented in Algorithm 1.

By Theorem 1 the probability of the event that at most k women are among the first i candidates in an unbiased environment, $Pr[X_i \leq k]$, is stored in $P[i, i-k]$ (defined in line 8 of Algorithm 1). Hence, we can verify the δ-fairness of a given ranking using Algorithm 2.

2.2 Obtaining Fair Ranking With Highest Utility

The main goal of Algorithm 3 is to find a fair ranking that has the maximum utility among all possible fair rankings. In order to do so, we follow a sequence of greedy operations and use dynamic programming to choose the best (highest utility) ranking at each

Algorithm 1 Probabilities Of Partial Sets

Input Dataset of candidates C.
Let n, n_1 and n_2 be the total number of candidates, women, and men respectively.
Let Q be a $n \times n_2$ matrix in which the $Q[i, m]$ corresponds to the probability of the event that m men occur in the first i candidates of an unbiased ranking.
Initialize Q:

$$Q[i,m] = \begin{cases} 0 & : \ i < m \\ \frac{n_1}{n} & : \ (i,m) = (1,0) \\ \frac{n_2}{n} & : \ (i,m) = (1,1) \end{cases}$$

for $i = 1, 2, \ldots n$ **do**
  for $m = 1, 2, \ldots, \min(i, n_2)$ **do**

$$Q\,[i,m] = \left(\frac{n_2 - (m-1)}{n - (i-1)}\right) Q\,[i-1, m-1] + \left(\frac{n_1 - (i-1-m)}{n - (i-1)}\right) Q\,[i-1, m]$$

  end for
end for
Let P be a $n \times n_2$ matrix in which the $P[i, m]$ corresponds to the probability of the event that at least m men are among the first i candidates of an unbiased ranking.
Initialize P:

$$P[i,m] = \begin{cases} 0 & : \ i < m \\ 1 & : \ m = 0 \\ \frac{n_2}{n} & : \ (i,m) = (1,1) \end{cases}$$

for $i = 1, 2, \ldots n$ **do**
  for $m = 1, 2, \ldots, \min(i, n_2)$ **do**

$$P\,[i,m] = \sum_{j=m}^{i} Q[i,j]$$

  end for
end for
return P

Algorithm 2 VerifyFairnessByPartialSets

Input Dataset of candidates C, a permutation (ranking) function τ, a real positive number δ
Let n be the total number of candidates.
$P \leftarrow$ Probabilities Of Partial Sets(C)
for $1 \le i \le n$ **do**:
m_i = number of men in $\{\tau(1), \ldots, \tau(i)\}$
  if $P[i, m_i] < \delta$ **then** report **unfair** and terminate.
  end if
end for
Report **fair**.

step. The following theorem allows us to construct the required ranking permutation inductively, as the algorithm 3 does.

Theorem 2. *For a positive real number δ and a candidate set C, Algorithm 3 outputs a δ-fair ranking that has the highest utility.*

Algorithm 3 FindTheBestRankingByPsets

Input candidate set C, observed scores of candidates $\hat{Y}$, a real positive number δ
Let n, n_1 and n_2 be the total number of candidates, women, and men respectively.
Let Y_w and Y_m be the sorted scores of women and men, respectively.
$P \leftarrow$ CalculateProbabilitiesOfPartialSets(C) ▷ Algorithm 1
Let U be an $n \times n_2$ matrix that stores the highest utility that can be obtained by a fair ranking of i candidates with m men among them in $U\,[i, m]$.
Let R be a table with lists as entries that store the corresponding ranking of $U\,[i, m]$.
Initialize U as follows:

$$U\,[1,0] = \begin{cases} Y_w\,[0] & \text{If } P\,[1,0] > \delta \\ -\infty & \text{O.W} \end{cases}$$

$$U\,[1,1] = \begin{cases} Y_m\,[0] & \text{If } P\,[1,1] > \delta \\ -\infty & \text{O.W} \end{cases}$$

for $i = 2, \ldots, n$ **do**
  for $m = 0, \ldots, \min(i, n_2)$ **do**
    if $P[i, m] < \delta$ **then** $U[i, m] = -\infty$, break
    end if
    if $i - m - 1 < n_1$ **then**

$$u_1 = \frac{Y_w[i-m-1]}{\log_2(i+1)} + U\,[i-1, m]$$

    end if
    if m > 0 **then**

$$u_2 = \frac{Y_m[m-1]}{\log_2(i+1)} + U\,[i-1, m-1]$$

    end if
    Handle the extreme cases of $i - 1 - m = n_1$ and $m = 0$.
    $U[i, m] = \max(u_1, u_2)$
    Update R
  end for
end for
Let $\pi = R[n, n_2]$.
Output π

3 EXPERIMENTAL RESULTS

In this section, we report the experimental results in which we compared the average *true utility* of several algorithms on several synthetic data sets. Synthetic datasets are artificially created datasets that imitate the properties and structure of real-world data through a clear and understandable process. By true utility we mean the value of the utility function (that is introduced in (1)) on the unbiased scores which in reality we are not aware of, but since we are using synthetic datasets, we can assume that the unbiased scores are provided initially. Each data set consists of a set of candidates which are grouped by their gender and a set of unbiased scores for all of them which comes from a distribution independent

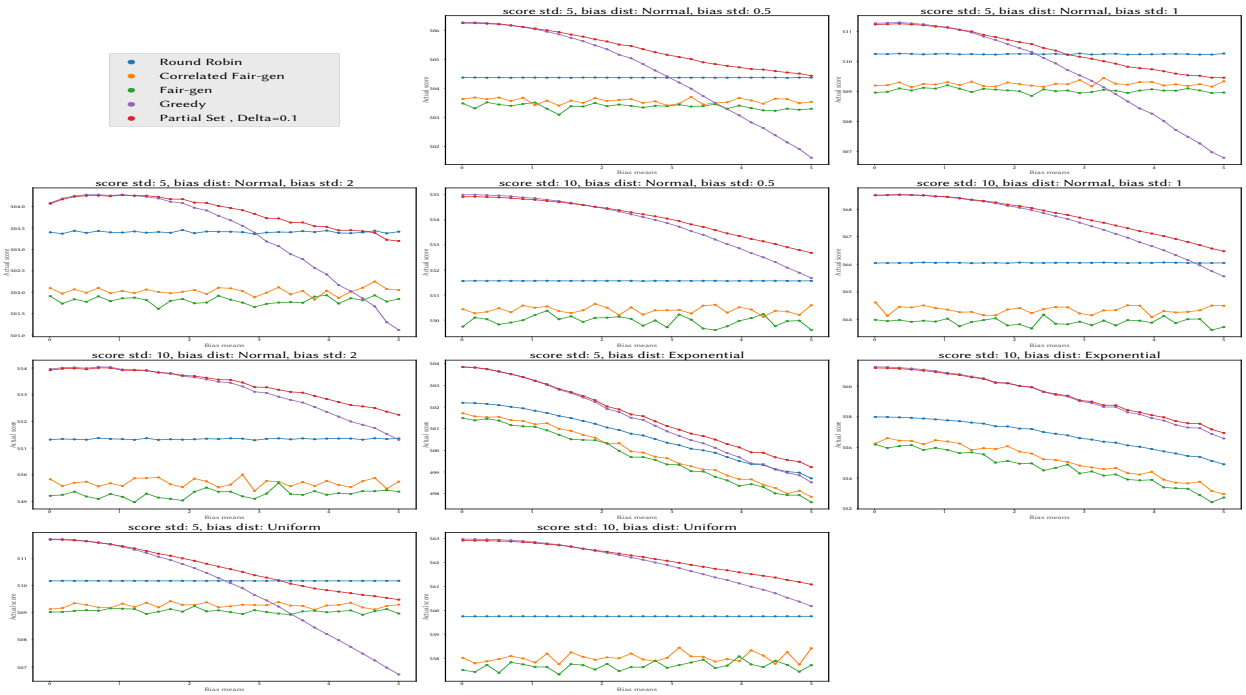

Figure 1: Utility of Algorithms Over Unbiased Scores, 50 women - 50 men

of their gender. As the literature implies, we assume the male candidates are the privileged ones so we set the observed scores of the male candidates the same as their unbiased scores. The observed scores of women candidates are obtained by their unbiased scores decremented by a random bias. We assume the unbiased scores come from a normal distribution and without loss of generality[1], we set the average to be 15. We report the results for two different standard deviations 5 and 10. The distribution of bias may vary, but here because of the paper size limit, we just study two cases of Normal and Uniform and Exponential distribution. For the case that the bias comes from Normal or Exponential distribution, we report the results for the bias average range of 0 to 5 and for the Normal case specifically, we report the outcome for three different standard deviations 0.5, 1, and 2 which seem more realistic in practice. In the body of the paper, we study the cases where the number of male and female candidates are equal. In the appendix, we provide the experiments where the portion of men and women are not equal.

We implemented Algorithm 3 (which is called *Partial Set* in our graphs) as well as some other algorithms motivated by the literature. Here we re-introduce some previously studied algorithms that will be used in our experiments.

Round Robin: This is the most trivial approach for satisfying fairness criteria. If the portion of men to women is β, we simply put the best-unranked woman after each β men. For the sake of simplicity, we suppose β is 1, $\frac{1}{3}$ and 3.

[1]Because by the linearity of expectation, if we add a constant value to all of the scores, the mean of the scores would be shifted by the same value. Moreover, this constant shift does not change the order of the candidates and the utility function as defined in (1) would be shifted by a function of that constant value.

Correlated Fair Gen: Yang et al. [2] propose an algorithm (called Ranking Generator) which randomly ensures that the number of protected candidates does not fall far below the minority proportion p. For each step of the Ranking generator algorithm, a Bernoulli experiment with the success probability $1-p$ is done and if the experiment succeeds, we put a man in that place.

Correlated Fair Gen: In the correlated approach, called *Correlated Fair Gen*, we update the minority proportion p in each step and place the candidate using a Bernoulli experiment as in the above algorithm. This modified version is represented here just to enrich our experiments.

Due to the page limit, we just report the experiments of the cases in which the population of men and the population of women are equal. We note that, we do not observe a huge change in the behavior of the algorithms when we change the portion of men and women.

The main statement that we want to conclude from these experiments is that the Partial Set algorithm, which is in some sense more moderate than Greedy and Round Robin, *almost* every time can do better than both of them on unbiased scores. Because the Partial Set algorithm cares about fairness and utility at the same time and is a mix of Greedy and Round Robin reasonably. In the following experiments, the comparisons clarify when our algorithm does and when it does not better than the other two.

As shown in Figure 1, in all of the experiments, Partial Set and Greedy algorithms are almost the same when the bias average is low. And when the bias average increases, Greedy goes down faster than any other algorithm and if the bias average is not higher

than a large value, the Partial Set algorithm has the highest utility among them all. But when the bias average exceeds a certain threshold, the Round Robin wins and it makes sense.

REFERENCES

[1] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM* 64, 4 (2021), 136–143.

[2] 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.

[3] Meike Zehlike, Tom Sühr, Ricardo Baeza-Yates, Francesco Bonchi, Carlos Castillo, and Sara Hajian. 2022. Fair Top-k Ranking with multiple protected groups. *Information processing & management* 59, 1 (2022), 102707.

[4] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2021. Fairness in ranking: A survey. *arXiv preprint arXiv:2103.14000* (2021).