Fairness in Ranking under Uncertainty

Anonymous Author(s)
Affiliation
Address
email

Abstract

Fairness has emerged as an important consideration in algorithmic decision making. Unfairness occurs when an agent with higher merit obtains a worse outcome than an agent with lower merit. Our central point is that a primary cause of unfairness is uncertainty. A principal or algorithm making decisions never has access to the agents' true merit, and instead uses proxy features that only imperfectly predict merit (e.g., GPA, star ratings, recommendation letters). None of these ever fully capture an agent's merit; yet existing approaches have mostly been defining fairness notions directly based on observed features and outcomes.

Our primary point is that it is more principled to acknowledge and model the uncertainty explicitly. The role of observed features is to give rise to a posterior distribution of the agents' merits. We use this viewpoint to define a notion of approximate fairness in ranking. We call an algorithm ϕ-fair (for $\phi \in [0, 1]$) if it has the following property for all agents x and all k: if agent x is among the top k agents with respect to *merit* with probability at least ρ (according to the posterior merit distribution), then the algorithm places the agent among the top k agents in its *ranking* with probability at least $\phi\rho$.

We show how to compute,rankings that optimally trade off approximate fairness against utility to the principal. In addition to the theoretical characterization, we present an empirical analysis of the potential impact of the approach in simulation studies. For real-world validation, we applied the approach in the context of a paper recommendation system that we built and fielded at a large conference.

1 Introduction

Fairness is an important consideration in decision-making, in particular when a limited resource must be allocated among multiple agents by a principal (or decision maker). A widely accepted tenet of fairness is that if an agent B does not have stronger merits for the resource than A, then B should not get more of the resource than A; *merit* could be a qualification (e.g., job performance), a need (e.g., disaster relief), or some other measure of eligibility.

The motivation for our work is that uncertainty about merits is a primary reason that a principal's allocations can violate this tenet and thereby lead to unfair outcomes. Were agents' merits fully observable, it would be both fair and in the principal's best interest to rank agents by their merit. However, actual merits are practically always unobservable. Consider the following standard algorithmic decision making environments: (1) An e-commerce or recommender platform (the principal) displays items (the agents) in response to a user query. An item's merit is the utility the user would derive from it, whereas the platform can only observe numerical ratings, text reviews, the user's past history, and similar features. (2) A job recommendation site or employer (the principal) wants to recommend/hire one or more applicants (the agents). The merit of an applicant is her (future) performance on the job over a period of time, whereas the site or employer can only observe (past) grades, test scores, recommendation letters, performance in interviews, and similar assessments.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

In both of these examples — and essentially all others in which algorithms are called upon to make allocation decisions between agents — uncertainty about merit is unavoidable, and arises from multiple sources: (1) the training data of a machine learning algorithms is a random sample, (2) the features themselves often come from a random process, and (3) the merit itself may only be revealed in the future after a random process (e.g., whether an item is sold or an employee performs well). Given that decisions will be made in the presence of uncertainty, it is important to define the notion of *fairness* under uncertainty. Extending the aforementioned tenet that "if agent B has less merit than A, then B should not be treated better than A," we state the following generalization to uncertainty about merit, first for just two agents:

Axiom 1. If A has merit greater than or equal to B with probability at least ρ, then a fair policy should treat A at least as well as B with probability at least ρ.

This being an axiom, we cannot offer a mathematical justification. It captures an inherent sense of fairness in the absence of enough information, and it converges to the conventional tenet as uncertainty is reduced. In particular, consider the following situation: two items A, B with 10 reviews each have average star ratings of 3.9 and 3.8, respectively; or two job applicants A, B have GPAs of 3.9 and 3.8. While this constitutes some (weak) evidence that A may have more merit than B, this evidence is very uncertain. The posterior merit distributions based on the available information should reflect this uncertainty by having non-trivial variance; our axiom then implies that A and B must be treated similarly to achieve fairness. In particular, it would be highly unfair to *deterministically* rank A ahead of B (or vice versa). Our main point is that this uncertainty, rather than the specific numerical values of 3.9 and 3.8, is what should make a mechanism treat A and B similarly.

1.1 Our Contributions

We study fairness in the presence of uncertainty specifically for the generalization where the principal must rank n items. Our main contribution is the fairness framework, giving definitions of fairness in ranking in the presence of uncertainty. This framework, including extensions to approximate notions of fairness, is presented in Section § 2. We believe that uncertainty of merit is one of the most important sources of unfairness, and modeling it explicitly and axiomatically is key to addressing it.

Next, in Section § 3, we present algorithms for a principal to achieve (approximately) fair ranking distributions. A simple way is for the principal to mix between an optimal (unfair) ranking and (perfectly fair) Thompson sampling. This policy is not optimal for the principal's utility, and we present an efficient LP-based algorithm that achieves an optimal ranking distribution for the principal, subject to an approximate fairness constraint.

We next explore empirically to what extent a focus on fairness towards the agents reduces the principal's utility. We do so with two extensive sets of experiments: one described in Section § 4 on existing data, and one described in Section § 5 "in the wild." In the first set of experiments, we consider movie recommendations based on the standard MovieLens dataset and investigate to what extent fairness towards movies would result in lower utility for users of the system. The second experiment was carried out at a scientific conference, where we implemented and fielded a paper recommendation system, and half of the conference attendees using the system received rankings that were modified to ensure greater fairness towards papers. We report on various metrics that capture the level of engagement of conference participants based on which group they were assigned to.

The upshot of our experiments and theoretical analysis is that at least in the settings we have studied, high levels of fairness can be achieved at a relatively small loss in utility for the principal and the system's users.

2 Ranking with Uncertain Merits

We are interested in ranking policies for a principal (the ranking system, such as an e-commerce platform or a job portal in our earlier examples) whose goal is to rank a set $\mathcal{X}$ of n agents (such as products or applicants). The principal observes some evidence for the merit of the agents, and must produce a distribution over rankings trading off fairness to the agents against the principal's utility. For the agents, a higher rank is always more desirable than a lower rank.

2.1 Rankings and Ranking Distributions

We use $\Sigma(\mathcal{X})$ to denote the set of all $n!$ rankings, and $\Pi(\mathcal{X})$ for the set of all distributions over $\Sigma(\mathcal{X})$. We express a ranking $\sigma \in \Sigma(\mathcal{X})$ in terms of the agents assigned to given positions, i.e., $\sigma(k)$ is the agent in position k. A ranking distribution $\pi \in \Pi(\mathcal{X})$ can be represented by the $n!$ probabilities $\pi(\sigma)$ of the rankings $\sigma \in \Sigma(\mathcal{X})$. However, all the information relevant for our purposes can be represented more compactly using the *Marginal Rank Distribution*: we write $p^{(\pi)}_{x,k} = \sum_{\sigma:\sigma(k)=x} \pi(\sigma)$ for the probability under π that agent $x \in \mathcal{X}$ is in position k in the ranking. We let $\mathcal{P}^{(\pi)} = (p^{(\pi)}_{x,k})_{x,k}$ denote the $n \times n$ matrix of all marginal rank probabilities.

$\mathcal{P}^{(\pi)}$ is doubly stochastic, i.e., the sum of each row and column is 1. While π uniquely defines $\mathcal{P}^{(\pi)}$, the converse mapping may not be unique. Given a doubly stochastic matrix $\mathcal{P}$, the Birkhoff-von Neumann decomposition (Birkhoff, 1946) can be used to compute *some* ranking distribution π consistent with $\mathcal{P}$, i.e., $\mathcal{P}^{(\pi)} = \mathcal{P}$; any consistent distribution π will suffice for our purposes.

2.2 Merit, Uncertainty, and Fairness

The principal must determine a distribution over rankings of the agents. This distribution will be based on some evidence for the agents' merits. This evidence could take the form of star ratings and reviews of products (combined with the site visitor's partially known preferences), or GPA, test scores, and recommendation letters of an applicant. Our main departure from past work on individual fairness (following (Dwork et al., 2012)) is that we do not view this evidence as having inherent meaning; rather, its sole role is to induce a posterior joint distribution over the agents' merits.

The merit of agent x is $v_x \in \mathbb{R}$, and we write $\boldsymbol{v} = (v_x)_{x\in\mathcal{X}}$ for the vector of all agents' merits. Based on all observed evidence, the principal can infer a distribution Γ over agents' merits using any suitable Bayesian inference procedure. Since the particular Bayesian model depends on the application, for our purposes, we merely assume that a posterior distribution Γ was inferred using best practices and that ideally, this model is open to verification and audit.

We write $\Gamma(\boldsymbol{v})$ for the probability of merits $\boldsymbol{v}$ under Γ. We emphasize that the distribution will typically not be independent over entries of $\boldsymbol{v}$ — for example, students' merit conditioned on observed grades will be correlated via common grade inflation if they took the same class. To avoid awkward tie-breaking issues, we assume that $v_x \neq v_y$ for all distinct $x, y \in \mathcal{X}$ and all $\boldsymbol{v}$ in the support of Γ. This side-steps having to define the notion of top-k lists with ties, and comes at little cost in expressivity, as any tie-breaking would typically be encoded in slight perturbations to the v_x anyway.

We write $\mathcal{M}^{(\boldsymbol{v})}_{x,k}$ for the event that under $\boldsymbol{v}$, agent x is among the top k agents with respect to merit, i.e., that $|\{x' \mid v_{x'} > v_x\}| < k$. We now come to our key definition of approximate fairness.

Definition 2.1 (Approximately Fair Ranking Distribution)**.** A ranking distribution π is *ϕ-fair* iff

$$\sum_{k'=1}^{k} p^{(\pi)}_{x,k'} \geq \phi \cdot \mathbb{P}_{\boldsymbol{v}\sim\Gamma}\left[\mathcal{M}^{(\boldsymbol{v})}_{x,k}\right] \tag{1}$$

for all agents x and positions k. That is, the ranking distribution π ranks x at position k or above with at least a ϕ fraction of the probability that x is actually among the top k agents according to Γ. Furthermore, π is *fair* iff it is 1-fair.

The reason for defining ϕ-approximately fair ranking distributions (rather than just fair distributions) is that fairness typically comes at a cost to the principal (such as lower expected clickthrough or lower expected performance of recommended employees). For example, if the v_x are probabilities that a user will purchase products on an e-commerce site, then deterministically ranking by decreasing $\mathbb{E}_\Gamma[v_x]$ is the principal's optimal ranking under common assumptions about user behavior; yet, being deterministic, it is highly unfair. Our definition of approximate fairness allows, e.g., a policymaker to choose a tradeoff regarding how much fairness (with resulting expected utility loss) to require from the principal. Notice that for $\phi = 0$, the principal is unconstrained.

We remark that in the definition of (approximate) fairness, the numerical values of v_x only matter insofar as comparison is concerned. A useful implication is the following: let $f : \mathbb{R} \to \mathbb{R}$ be any strictly monotone function. If we replace the distribution Γ with a distribution Γ' that draws the

vector $(f(v_x))_{x \in \mathcal{X}}$ with probability $\Gamma(\boldsymbol{v})$ (i.e., it replaces each entry v_x with $f(v_x)$), then if π was ϕ-fair with respect to Γ, it is also ϕ-fair with respect to Γ'.

2.3 The Principal's Utility

The principal's utility can be the profit of an e-commerce site or the satisfaction of its customers. We assume that the principal's utility for a ranking σ with agent merits $\boldsymbol{v}$ takes the form

$$U(\sigma \mid \boldsymbol{v}) = \sum_{k=1}^{n} w_k v_{\sigma(k)}, \tag{2}$$

where w_k is the position weight for position k in the ranking, and we assume that the w_k are non-increasing, i.e., the principal derives the most utility from earlier positions of the ranking.

The assumption that the utility from each position is factorable (i.e., of the form $w_k \cdot v_{\sigma(k)}$) is quite standard in the literature (Järvelin and Kekäläinen, 2002; Taylor et al., 2008). The assumption that the utility is *linear* in $v_{\sigma(k)}$ is in fact not restrictive at all. To see this, assume that the principal's utility were of the form $w_k \cdot f(v_{\sigma(k)})$ for some strictly increasing function f. As remarked above, the exact same fairness guarantees are achieved when the agents' merits v_x are replaced with $f(v_x)$; doing so preserves fairness, while in fact making the principal's utility linear in the merits. Some common examples of utility functions falling into this general framework are DCG with $w_k = 1/\log_2(1+k)$, Average Reciprocal Rank with $w_k = 1/k$, and Precision@K with $w_k = \mathbb{1}[k \leq K]/K$.

When the ranking and merits are drawn from distributions, the principal's utility is the expected utility under both sources of randomness:

$$U(\pi \mid \Gamma) = \mathbb{E}_{\sigma \sim \pi, \boldsymbol{v} \sim \Gamma} \left[U(\sigma \mid \boldsymbol{v}) \right]. \tag{3}$$

2.4 Discussion

Our definition is superficially similar to existing definitions of individual fairness (e.g., (Dwork et al., 2012; Joseph et al., 2016)), in that similar observable features often lead to similar outcomes. Importantly, though, it side-steps the need to define a similarity metric between agents in the feature space. Furthermore, it does not treat the observable attributes (such as star ratings) themselves as any notion of "merit." Instead, our central point is that agents' features should be viewed (solely) as noisy signals about the agents' merits and that a comparison of their merits — and the principal's uncertainty about the merits — should determine the agents' relative ranking. That moves the key task of quantifying individual fairness from articulating which features should be considered relevant for similarity, to articulating what inferences can be drawn about merit from observed features.

One may argue, rightfully, that from an operational perspective, this simply pushes the normative decisions into determining Γ. However, our main point is that this is where normative decisions indeed belong. Rather than trying to define ad hoc notions of fairness on observable features, the principal should determine what the features reveal (and don't reveal) about the agents' merits. In this context, it is worth remarking that the distribution Γ may also take care of systemic biases. For example, if an applicant's features like GPA are biased by particular characteristics (e.g., race or gender), Γ can/should reflect that applicants can have the same merit even if they do not have the same GPA. Again, the important distinction here is to not treat observable features such as GPA or star ratings as merit, but rather as noisy and imperfect (and possibly biased) *signals* about merit.

An additional benefit of requiring the use of fair ranking policies, discussed in depth in Appendix § D, is that it makes the principal bear more of the cost of unfairness, and thereby incentivizes the principal to improve the distribution Γ. To see this at a high level, notice that if Γ precisely revealed merits, then the optimal and fair policies would coincide. In the presence of uncertainty, an unrestricted principal will optimize utility, and in particular do better than a principal who is constrained to be (partially or completely) fair. Thus, a fair principal stands to gain more by obtaining perfect information. The example we present in Appendix § D shows that this difference can be substantial, i.e., the information acquisition incentives for a fair principal can be much higher.

We emphasize here that we believe that low-stakes scenarios repeated often are typically a better fit for any randomized approach (including ours) than high-stakes one-shot games. The reasons include that for a high-stakes situation, a principal may be less willing to trade off utility for fairness and that

for a single decision, verifying that a decision was made probabilistically could be hard or impossible. Thus, ideal scenarios to keep in mind are e-commerce websites or other recommendation systems that repeatedly choose a ranking of products (or papers or other items) in response to a customer's query.

In terms of the existing literature, our work generally relates to existing works on addressing bias and fairness issues in algorithmic decision making, especially machine learning (Barocas and Selbst, 2016; Hardt et al., 2016; Dwork et al., 2012)(see Barocas et al. (2019) for a summary). More specifically, our work is close to recent work on defining fairness notions for rankings which can be categorized into three groups: composition-based notions (Yang and Stoyanovich, 2017; Celis et al., 2017; Zehlike et al., 2017), opportunity-based notions (Singh and Joachims, 2018, 2019; Biega et al., 2018), and evidence-based notions (Dwork et al., 2019a; Beutel et al., 2019; Narasimhan et al., 2020). Our fairness axiom combines the opportunity-based and evidence based notions by stating that the economic opportunity allocated to the agents must be consistent with the existing evidence about their relative ordering. A detailed discussion of this and other related work is provided in Appendix § A.

3 Optimal and Fair Policies

For a distribution Γ over merits, let σ^*_Γ be the ranking which sorts the agents by expected merit, i.e., by non-increasing $\mathbb{E}_{\boldsymbol{v}\sim\Gamma}[v_x]$. The following proposition follows because the position weights w_k are non-increasing; its proof, along with all other missing proofs, is given in Appendix § B.

Proposition 3.1. σ^*_Γ is a utility-maximizing ranking policy for the principal.

While this policy conforms to the Probability Ranking Principle (Robertson, 1977), it violates Axiom 1 for ranking fairness when merits are uncertain. We define a natural solution for a 1-fair ranking distribution based on Thompson Sampling:

Definition 3.1 (Thompson Sampling Ranking Distribution)**.** Define π^{TS}_Γ as follows: first, draw a vector of merits $\boldsymbol{v} \sim \Gamma$, then rank the agents by decreasing merits in $\boldsymbol{v}$.

That π^{TS}_Γ is 1-fair follows directly from the definition of fairness. By definition, it ranks each agent x in position k with exactly the probability that x has k-th highest merit.

Proposition 3.2. π^{TS}_Γ is a 1-fair ranking distribution.

3.1 Trading Off Utility and Fairness

One straightforward way of trading off between the two objectives of fairness and principal's utility is to randomize between the two policies π^{TS}_Γ and π^*.

Definition 3.2 (OPT/TS-Mixing)**.** The OPT/TS-Mixing ranking policy $\pi^{\mathrm{Mix},\phi}$ randomizes between π^{TS}_Γ and π^*_Γ with probabilities ϕ and $1-\phi$, respectively.

The following lemma gives guarantees for such randomization (but we will later see that this strategy is suboptimal).

Lemma 3.1. Consider two ranking policies π_1 and π_2 such that π_1 is ϕ_1-fair and π_2 is ϕ_2-fair. A policy that randomizes between π_1 and π_2 with probabilities q and $1-q$, respectively, is at least $(q\phi_1 + (1-q)\phi_2)$-fair and obtains expected utility $qU(\pi_1 \,|\, \Gamma) + (1-q)U(\pi_2 \,|\, \Gamma)$.

Corollary 3.1. The ranking policy $\pi^{\mathrm{Mix},\phi}$ is ϕ-fair.

By definition, $\pi^{\mathrm{Mix},\phi=0}$ has the highest utility among all 0-fair ranking policies. Furthermore, all 1-fair policies achieve the same utility since the fairness axiom for $\phi = 1$ completely determines the marginal rank probabilities (Lemma B.1 in Appendix § B). However, while $\pi^{\mathrm{Mix},\phi=0}$ and $\pi^{\mathrm{Mix},\phi=1}$ have the highest utility among 0-fair and 1-fair ranking policies, respectively, $\pi^{\mathrm{Mix},\phi}$ will typically not have maximum utility for the principal among all the ϕ-fair ranking policies for other values of $\phi \in (0,1)$. We illustrate this with a concrete example with $n = 3$ agents in Appendix § C.

3.2 Optimizing Utility for ϕ-Fair Rankings

We now formulate a linear program for computing the policy $\pi^{\mathrm{LP},\phi}$ that maximizes the principal's utility, subject to being ϕ-fair. The variables of the linear program are the marginal rank probabilities

$p^{(\pi)}_{x,k}$ of the distribution π to be determined. Then, by Equation (3) and linearity of expectation, the principal's expected utility can be written as $U(\pi \mid \Gamma) = \sum_{x \in \mathcal{X}} \sum_k p^{(\pi)}_{x,k} \cdot \mathbb{E}_{\boldsymbol{v} \sim \Gamma}[v_x] \cdot w_k$. We use this linear form of the utilities to write the optimization problem as the following LP with variables $p_{x,k}$ (omitting π from the notation):

$$\begin{array}{lll} \text{Maximize} & \sum_x \sum_k p_{x,k} \cdot \mathbb{E}_{\boldsymbol{v} \sim \Gamma}[v_x] \cdot w_k & \\ \text{subject to} & \sum_{k'=1}^{k} p_{x,k'} \geq \phi \cdot \mathbb{P}_{\boldsymbol{v} \sim \Gamma}\left[\mathcal{M}^{(\boldsymbol{v})}_{x,k}\right] & \text{for all } x, k \\ & \sum_{k=1}^{n} p_{x,k} = 1 & \text{for all } x \\ & \sum_{i=1}^{n} p_{x,k} = 1 & \text{for all } k \\ & 0 \leq p_{x,k} \leq 1 & \text{for all } x, k. \end{array} \tag{4}$$

In the LP, the first set of constraints captures ϕ-approximate fairness for all agents and positions, while the remaining constraints ensure that the marginal probabilities form a doubly stochastic matrix.

As a second step, the algorithm uses the Birkhoff-von Neumann (BvN) Decomposition of the matrix $\mathcal{P} = (p_{x,k})_{x,k}$ to explicitly obtain a distribution π over rankings such that π has marginals $p_{x,k}$. The Birkhoff-von Neumann Theorem (Birkhoff, 1946) states that the set of doubly stochastic matrices is the convex hull of the permutation matrices, which means that we can write $\mathcal{P} = \sum_\sigma q_\sigma \mathcal{P}^{(\sigma)}$, where $\mathcal{P}^{(\sigma)}$ is the binary permutation matrix corresponding to the deterministic ranking σ, and the q_σ form a probability distribution. It was already shown by Birkhoff (1946) how to find a polynomially sparse decomposition in polynomial time.

In order to solve the Linear Program (4), one needs to know $\mathbb{P}_{\boldsymbol{v} \sim \Gamma}\left[\mathcal{M}^{(\boldsymbol{v})}_{x,k}\right]$ for all i and k. For some distributions Γ (e.g., Example 1), these quantities can be calculated in closed form. For others, they can be estimated using Monte Carlo sampling. Such estimates may be off by small terms (which go to 0 as the number of samples grows large). This may result in a small loss of $\epsilon \to 0$ in the fairness guarantee which can be compensated by using a more aggressive fairness parameter $\phi' = \phi + \epsilon$ (when $\phi \leq 1 - \epsilon$). This translates the loss in fairness to a loss in utility, which also goes to 0 as the number of Monte Carlo samples grows.

4 Experimental Evaluation: MovieLens Dataset

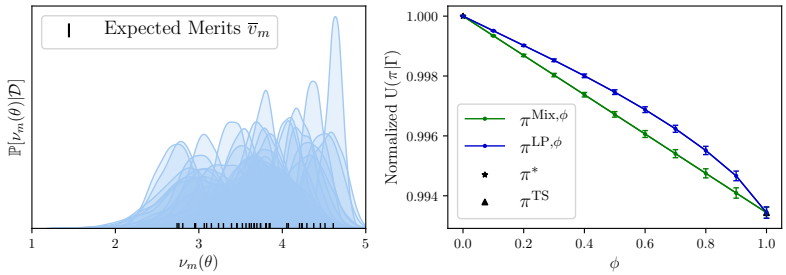
To evaluate our approach in a recommendation setting with a realistic preference distribution, we design the following experimental setup based on the MovieLens 100K (ML-100K) dataset. The dataset contains 100,000 ratings, by 600 users, on 9,000 movies belonging to 18 genres (Harper and Konstan, 2015). In our setup, for each user, the principal is a recommender system that has to generate a ranking of movies $\mathcal{S}_g$ for one of the genres g (e.g., Horror, Romance, Comedy) according to a notion of *merit* of the movies we define as follows.

Modeling the Merit Distribution. We define the (unknown) merit v_m of a movie m as the average rating of the movie across the user population[1] — this merit is unknown because most users have not seen/rated most movies. To be able to estimate this merit based on ratings in the ML-100K dataset, and to concretely define its underlying distribution and the corresponding fairness criteria, we define a generative model of user ratings. The model assumes that the rating of a movie $m \in \mathcal{S}_g$ is drawn from a multinomial distribution over $\{1, 2, 3, 4, 5\}$ with parameters $\boldsymbol{\theta}_m = (\theta_{m,1}, \ldots, \theta_{m,5})$. Assuming a Dirichlet prior, we can infer the posterior distribution of the ratings (and hence the merit) from the dataset in closed form, using merely the counts of each rating for each movie in the dataset (see Appendix § E.1.1 for details).

Based on this definition of merit and its uncertainty, we define and compare different ranking policies for this experimental setup as follows.

Utility Maximizing Ranking (π^*): We use the DCG function (Burges et al., 2005) with position weights $w_k = 1/\log_2(1+k)$ as our utility measure. Since the weights are indeed strictly decreasing, as described in Section § 3, the optimal ranking policy π^* sorts the movies (for the particular query)

[1] For a personalized ranking application, an alternative would be to choose each user's (mostly unknown) rating as the merit criterion instead of the average rating across the user population.



(a) (b)

Figure 1: (a) Posterior distribution of ratings (merit) for a subset of "Comedy" movies, (b) Tradeoff between Utility and Fairness, as captured by ϕ.

by decreasing expected merit, which is the expected average rating $\overline{v}_m \triangleq \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{P}[\boldsymbol{\theta}_m \mid \mathcal{D}]}\left[v_m(\boldsymbol{\theta})\right]$ under the posterior Dirichlet distribution in our case; here, $v_m(\boldsymbol{\theta})$ is the average rating of movie m corresponding to the parameter vector $\boldsymbol{\theta}$ sampled from the posterior.

Fully Fair Ranking Policy (π^{TS}): A fair ranking in this case ensures that, for all positions k, a movie is placed in the top k positions according to the posterior merit distribution. In this setup, a fully fair ranking policy π^{TS} is obtained by sampling the multinomial parameters $\boldsymbol{\theta}_m$ for each movie $m \in \mathcal{S}_g$ and computing $v_m(\boldsymbol{\theta}_m)$ to rank them:

$$\pi^{\mathrm{TS}}(\mathcal{S}_g) \sim \mathrm{argsort}_m\, v_m(\boldsymbol{\theta}_m) \text{ s.t. } \boldsymbol{\theta}_m \sim \mathbb{P}[\boldsymbol{\theta}_m|\mathcal{D}].$$

LP Ranking Policy ($\pi^{\mathrm{LP},\phi}$): The ϕ-fair policies $\pi^{\mathrm{LP},\phi}$ require the principal to have access to the probabilities $\mathbb{P}_{\boldsymbol{v}\sim\Gamma}\left[\mathcal{M}_{m,k}^{(\boldsymbol{v})}\right]$ which we estimate using $5 \cdot 10^4$ Monte Carlo samples, so that any estimation error becomes negligible.

OPT/TS-Mixing Ranking Policy ($\pi^{\mathrm{Mix},\phi}$): The policies $\pi^{\mathrm{Mix},\phi}$ randomize between the fully fair and utility-maximizing ranking policies with probabilities ϕ and $1-\phi$, respectively.

Observations and Results. In the experiments presented, we used the ranking policies π^*, π^{TS}, $\pi^{\mathrm{Mix},\phi}$ and $\pi^{\mathrm{LP},\phi}$ to create separate rankings for each of the 18 genres. For each genre, the task is to rank a random subset of 40 movies from that genre. To get a posterior with an interesting degree of uncertainty, we take a 10% i.i.d. samples from $\mathcal{D}$ to infer the posterior for each movie. We observe that the results are qualitatively consistent across genres, and we thus focus on detailed results for the genre "Comedy" as a representative example. Its posterior merit distribution over a subset is visualized in Figure 1(a). Note that substantial overlap exists between the marginal merit distributions of the movies, indicating that as opposed to π^* (which sorts based on the expected merits), the policy π^{TS} will randomize over many different rankings.

Observation 1: We evaluate the cost of fairness to the principal in terms of loss in utility, as well as the ability of $\pi^{\mathrm{LP},\phi}$ to minimize this cost for ϕ-fair rankings. Figure 1(b) shows this cost in terms of expected Normalized DCG (i.e., NDCG $= {}^{\mathrm{DCG}}/_{\max(\mathrm{DCG})}$ as in (Järvelin and Kekäläinen, 2002)). These results are averaged over 20 runs with different subsets of movies and different training samples. The leftmost end corresponds to the NDCG of π^*, while the rightmost point corresponds to the NDCG of the 1-fair policy π^{TS}.

The drop in NDCG is below one percentage point, which is consistent with the results for the other genres. We also conducted experiments with other values of s, data set sizes, and choices of w_k; even under the most extreme conditions, the drop was at most 2 percent. While this rather small drop may be surprising at first, we point out that uncertainty in the estimates affects the utility of both π^* and π^{TS}. By industry standards, a 2% drop in NDCG is considered quite substantial; however, it is not catastrophic and hence bodes well for possible adoption.

Observation 2: Figure 1(b) also compares the trade-off in NDCG in response to the fairness approximation parameter ϕ for both $\pi^{\mathrm{Mix},\phi}$ and $\pi^{\mathrm{LP},\phi}$. We observe that the utility-optimal policy $\pi^{\mathrm{LP},\phi}$ provides gains over $\pi^{\mathrm{Mix},\phi}$, especially for large values of ϕ. Thus, using $\pi^{\mathrm{LP},\phi}$ can further reduce the cost of fairness discussed above.

Observation 3: To provide intuition about the difference between $\pi^{\mathrm{Mix},\phi}$ and $\pi^{\mathrm{LP},\phi}$, Figure 4 in Appendix § E.1.3 visualizes the marginal rank distributions $p_{m,k}$, i.e., the probability that movie m is ranked at position k. The key distinction is that, for intermediate values of ϕ, $\pi^{\mathrm{LP},\phi}$ exploits a non-linear structure in the ranking distribution (to achieve a better trade-off) while $\pi^{\mathrm{Mix},\phi}$ merely interpolates linearly between the solutions for $\phi = 0$ and $\phi = 1$.

Based on these observations, in general, the utility loss due to fairness is small, and can be further reduced by optimizing the ranking distribution with the LP-based approach.These results are based on the definition of merit as the average rating of movies over the entire user population. A more realistic setting would personalize rankings for each user, with merit defined as the expected relevance of a movie to the user. In our experiments, the under such a setup were quite similar, and are hence omitted for brevity and clearer illustration.

5 Real-World Experiment: Paper Recommendation

To study the effect of deploying a fair ranking policy in a real ranking system, we built and fielded a paper recommendation system at a scientific conference. The goal of the experiment is to understand the impact of fairness under real user behavior, as opposed to simulated user behavior that is subject to modeling assumptions. Specifically, we seek to answer two questions: (a) Does a fair ranking policy lead to a more equitable distribution of exposure among the papers? (b) Does fairness substantially reduce the utility of the system to the users?

The users of the paper recommendation system were the participants of a virtual conference held in 2020. Signup and usage of the system was voluntary. Each user was recommended a personalized ranking of the papers published at the conference. This ranking was produced either by σ^* or by π^{TS}, and the assignment of users to treatment (π^{TS}) or control (σ^*) was randomized.

Modeling the Merit Distribution. The merit of a paper for a particular user is based on a relevance score $\mu_{u,i}$ that relates features of the user (e.g., bag-of-words representation of recent publications, co-authorship) to features of each conference paper (e.g., bag-of-words representation of paper, citations). Most prominently, the relevance score $\mu_{u,i}$ contains the TFIDF-weighted cosine similarity between the bag-of-words representations.

To model the uncertainty in $\mu_{u,i}$, we make the assumption that since all papers were accepted to the conference, they must have been deemed relevant to at least some fraction of the audience by the peer reviewers. Hence, papers with uniformly low relevance scores $\mu_{u,i}$ across users (e.g., ones introducing new research directions or bringing in novel techniques) must have a higher uncertainty in their relevance score. This assumption leads us to define the uncertainty as a normal distribution around the mean $\mu_{u,i}$ with standard deviation δ_i defined such that, for paper i, there exists at least one user who finds the paper highly relevant to their interests with high probability. (See Appendix § E.2.2 for details on how δ_i is calculated.)

Ranking Policies. Users in the control group $\mathcal{U}_{\pi^*}$ received rankings in decreasing order of $\mu_{u,i}$. Users in the treatment group $\mathcal{U}_{\pi^{\mathrm{TS}}}$ received rankings from the fair policy that sampled scores from the uncertainty distribution, $\hat{\mu}_{u,i} \sim \mathcal{N}(\mu_{u,i}, \delta_i)$, and ranked the papers by decreasing $\hat{\mu}_{u,i}$.

Results and Observations. We first analyze if the fair policy provided more equitable exposure to the papers. In this real-world evaluation, exposure is not equivalent to rank, but depends on whether users actually browsed to a given position in the ranking. Users could browse their ranking in pages of 5 recommendations each; we count a paper as *exposed* if the user scrolled to its page.

Observation 1: Figure 2 compares the histogram of exposure of the papers in the treatment and control groups. Under the fair policy, the number of papers in the lowest tier of exposure is roughly halved compared to the control condition. This verifies that the fair policy does have an impact on exposure in a real-world setting, and it aligns with our motivation that a fair ranking policy distributes exposure more equally among the ranked agents. This is also supported by comparing the Gini inequality coefficient (Gini, 1936) for the two distributions: $G(\pi^*) = 0.3302$, while $G(\pi^{\mathrm{TS}}) = 0.2996$ (where a smaller coefficient means less inequality in the distribution).

Observation 2: To evaluate the impact of fairness on user engagement, we analyze a range of engagement metrics as summarized in Table 1. While a total of 923 users signed up for the system

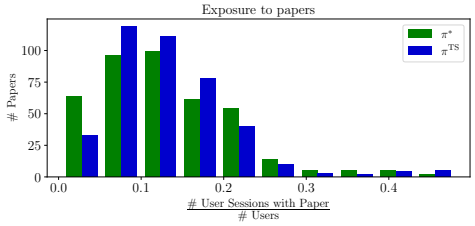


Figure 2: Distribution of the exposure of papers in treatment and control.

	Number of Users with activity		Average Activity Per User	
	π^*	π^{TS}	π^*	π^{TS}
Total Number of users	213	248	-	-
Num. of pages examined	-	-	10.8075	10.7984
`Read Abstract`	92	101	3.7230	2.6774
`Add to Calendar`	51	50	1.4366	0.8508
`Read PDF`	40	52	0.5258	0.5323
`Add Bookmark`	16	13	0.3192	0.6129

Table 1: User engagement under the two conditions π^* and π^{TS}. None of the differences are statistically significant. (For user actions, this is specifically due to the small sample size).

ahead of the conference (and were randomized into treatment and control groups), 462 never came to the system after all. Of the users that came, 213 users were in $\mathcal{U}_{\pi^*}$, and 248 users were in $\mathcal{U}_{\pi^{\mathrm{TS}}}$. Note that this difference is not caused by the treatment assignment, since users had no information about their assignment/ranking before entering the system. The first engagement metric we computed is the average number of pages that users viewed under both conditions. With roughly 10.8 pages (about 54 papers), engagement under both conditions was almost identical. Users also had other options to engage, but there is no clear difference between the conditions, either. On average, they read more paper abstracts and added more papers to their calendar under the control condition, but read more PDF and added more bookmarks under the treatment condition. However, none of these differences is significant at the 95% level for either a Mann-Whitney U test or a two sample t-test. While the sample size is small, these findings align with the findings on the synthetic data, namely that fairness did not appear to place a large cost on the principal (here representing the users).

6 Conclusions and Future Work

We believe that the focus on uncertainty we proposed in this paper constitutes a principled approach to capturing the intuitive notion of fairness to agents with similar features: rather than focusing on the features themselves, the key insight is that the features' similarity entails significant statistical uncertainty about which agent has more merit. Randomization provides a way to fight fire with fire, and axiomatize fairness in the presence of such uncertainty.

Our work raises a wealth of questions for future work. Perhaps most importantly, as discussed in Section § 2.4, to operationalize our proposed notion of fairness, it is important to derive principled merit distributions Γ based on the observed features. Our experiments were based on "reasonable" notions of merit distributions and concluded that fairness might not have to be very expensive to achieve for a principal. However, much more experimental work is needed to truly evaluate the impact of fair policies on the utility that is achieved. It would be particularly intriguing to investigate which types of real-world settings lend themselves to implementing fairness at little cost, and which force a steep trade-off between the two objectives.

Our work also raises several interesting theoretical questions. In Appendix § D, we show one setting in which forcing the principal to use a fair policy drastically increases the principal's incentives to form a more accurate posterior Γ. We did not prove a general result in this vein. We ask: will the incentives of a principal to learn a better posterior Γ always (weakly) increase if the principal is forced to be fairer? If true, tbis would provide a fascinating additional benefit of fairness requirements.

Another interesting question concerns the utility loss incurred by using the policy OPT/TS-Mixing. As we saw in Section § 3.1, OPT/TS-Mixing is in general not optimal. However, in the example from Section § 3.1 as well as in our experiments in Section § 4, the loss in utility was quite small. An interesting question would be to bound the worst-case loss in the utility of OPT/TS-Mixing, compared to the LP-based policy. In particular, this question is of interest due to the simplicity of the OPT/TS-Mixing policy; it does not require the computationally expensive solution of an LP or an explicit estimate of marginal rank probabilities under Γ.

References

Asudehy, A., Jagadishy, H., Stoyanovichz, J., and Das, G. (2017). Designing fair ranking schemes. *arXiv preprint arXiv:1712.09752*.

Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org. `http://www.fairmlbook.org`.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *Cal. L. Rev.*

Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E. H., et al. (2019). Fairness in recommendation ranking through pairwise comparisons. In *KDD*.

Biega, A. J., Gummadi, K. P., and Weikum, G. (2018). Equity of attention: Amortizing individual fairness in rankings. In *SIGIR*.

Birkhoff, G. (1946). Tres observaciones sobre el algebra lineal. *Univ. Nac. Tucuman*.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., and Hullender, G. (2005). Learning to rank using gradient descent. In *ICML*, pages 89–96.

Calders, T., Kamiran, F., and Pechenizkiy, M. (2009). Building classifiers with independency constraints. In *Data mining workshops, ICDMW*, pages 13–18.

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *NIPS*.

Carbonell, J. and Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*.

Celis, L. E., Straszak, D., and Vishnoi, N. K. (2017). Ranking with fairness constraints. *arXiv preprint arXiv:1704.06840*.

Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. In *SIGIR*.

Diaz, F., Mitra, B., Ekstrand, M. D., Biega, A. J., and Carterette, B. (2020). Evaluating stochastic rankings with expected exposure. In *CIKM*.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *ITCS*. ACM.

Dwork, C., Kim, M. P., Reingold, O., Rothblum, G. N., and Yona, G. (2019a). Learning from outcomes: Evidence-based rankings. *auditing*, 14:18.

Dwork, C., Kim, M. P., Reingold, O., Rothblum, G. N., and Yona, G. (2019b). Learning from outcomes: Evidence-based rankings. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE.

Ghosh, A., Dutt, R., and Wilson, C. (2021). When fair ranking meets uncertain inference. *arXiv preprint arXiv:2105.02091*.

Gini, C. (1936). On the measure of concentration with special reference to income and statistics. *Colorado College Publication, General Series*.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *NIPS*.

Harper, F. M. and Konstan, J. A. (2015). The movielens datasets: History and context. *ACM TIIS*.

Heidari, H. and Krause, A. (2018). Preventing disparate treatment in sequential decision making. In *IJCAI*.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*.

Joseph, M., Kearns, M., Morgenstern, J., and Roth, A. (2016). Fairness in learning: Classic and contextual bandits. In *NIPS*.

Kallus, N. and Zhou, A. (2019). The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. *NeurIPS*.

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In *NIPS*.

Kim, M. P., Ghorbani, A., and Zou, J. (2019). Multiaccuracy: Black-box post-processing for fairness in classification. In *AAAI Conference on AI, Ethics, and Society*.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *NIPS*.

Lahoti, P., Gummadi, K. P., and Weikum, G. (2019). Operationalizing individual fairness with pairwise fair representations. *arXiv preprint arXiv:1907.01439*.

Liu, Y., Radanovic, G., Dimitrakakis, C., Mandal, D., and Parkes, D. C. (2017). Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875*.

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

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Mehrotra, R., McInerney, J., Bouchard, H., Lalmas, M., and Diaz, F. (2018). Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *CIKM*.

Morik, M., Singh, A., Hong, J., and Joachims, T. (2020). Controlling fairness and bias in dynamic learning-to-rank. In *SIGIR*.

Narasimhan, H., Cotter, A., Gupta, M., and Wang, S. (2020). Pairwise fairness for ranking and regression. In *AAAI*.

Patil, V., Ghalme, G., Nair, V., and Narahari, Y. (2020). Achieving fairness in the stochastic multi-armed bandit problem. In *AAAI*.

Penha, G. and Hauff, C. (2021). On the calibration and uncertainty of neural learning to rank models. *arXiv preprint arXiv:2101.04356*.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On fairness and calibration. *NIPS*.

Prost, F., Awasthi, P., Blumm, N., Kumthekar, A., Potter, T., Wei, L., Wang, X., Chi, E. H., Chen, J., and Beutel, A. (2021). Measuring model fairness under noisy covariates: A theoretical perspective. *arXiv preprint arXiv:2105.09985*.

Radlinski, F., Bennett, P. N., Carterette, B., and Joachims, T. (2009). Redundancy, diversity and interdependent document relevance. In *ACM SIGIR Forum*.

Robertson, S. E. (1977). The probability ranking principle in ir. *Journal of documentation*.

Schumann, C., Lang, Z., Mattei, N., and Dickerson, J. P. (2019). Group fairness in bandit arm selection. *arXiv preprint arXiv:1912.03802*.

Singh, A. and Joachims, T. (2018). Fairness of exposure in rankings. In *KDD*.

Singh, A. and Joachims, T. (2019). Policy learning for fairness in ranking. In *NeurIPS*.

Soufiani, H. A., Parkes, D. C., and Xia, L. (2012). Random utility theory for social choice. In *NIPS*.

Taylor, M., Guiver, J., Robertson, S., and Minka, T. (2008). Softrank: optimizing non-smooth rank metrics. In *WSDM*.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. (2017). Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*.

Yang, K. and Stoyanovich, J. (2017). Measuring fairness in ranked outputs. *SSDBM*.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *WWW*.

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., and Baeza-Yates, R. (2017). Fa* ir: A fair top-k ranking algorithm. *CIKM*.

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

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *ICML*.

Zliobaite, I. (2015). On the relation between accuracy and fairness in binary classification. *FATML Workshop at ICML*.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We discuss them in Section § 6 and Section § 2.4.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Our work focuses on making algorithmic decision making systems more fair. To motivate our work, we indeed start by discussing the negative impacts of existing algorithms that do not consider uncertainty in decision making.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Some proofs were moved to the Supplementary Material due to space restrictions. See Appendix Section § B.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Uploaded with the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Figure 3.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] Our code runs in less than an hour on a single-core CPU with enough memory to load the entire Movielens Dataset. We do not anticipate this to be a challenge in reproduction of our results.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] The creators of the Movielens dataset are cited in Section § 4.
   (b) Did you mention the license of the assets? [Yes] The license is included with the dataset that can be downloaded from the URL provided with the code.

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A] The dataset is publicly available. A URL is provided with the code to download the dataset.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] We did not discuss the data collection process of the dataset in this paper but the original publication by Harper and Konstan (2015) provides detail about the process of data collection.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] The ML-100K dataset (Harper and Konstan, 2015) is anonymized, and does not contain any offensive content.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] The real-world experiments were performed in an academic conference held virtually in 2020. The details of the instructions and screenshots have been witheld to preserve anonymity of the authors.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] The protocol for our experiment was approved as "exempt from IRB review". We created a consent form for the users to read and agree to when signing up to use our system. We have not included the text of the form to preserve anonymity of the authors.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] The users agreed to use the system without any financial compensation.