Proceedings of Machine Learning Research 279:8–37, 2024 Algorithmic Fairness Through the Lens of Metrics and Evaluation

From Models to Systems: A Comprehensive Framework for AI System Fairness in Compositional Recommender Systems

Brian Hsu BHSU@LINKEDIN.COM
LinkedIn Corporation
Cyrus DiCiccio CJD48@CORNELL.EDU
Independent Researcher
Natesh S. Pillai NAPILLAI@LINKEDIN.COM
LinkedIn Corporation, Harvard University
Hongseok Namkoong NAMKOONG@GSB.COLUMBIA.EDU
LinkedIn Corporation, Columbia Business School

Editors: Miriam Rateike, Awa Dieng, Jamelle Watson-Daniels, Ferdinando Fioretto, Golnoosh Farnadi

Abstract

Fairness research in machine learning often centers on ensuring equitable performance of individual models. However, real-world recommendation systems are built on multiple models and even multiple stages, from candidate retrieval to scoring and serving, which raises challenges for responsible development and deployment. This *AI system-level view*, as highlighted by regulations like the EU AI Act, necessitates moving beyond auditing individual models as independent entities. We propose a holistic framework for modeling AI system-level fairness, focusing on the end-utility delivered to diverse user groups, and consider interactions between components such as retrieval and scoring models. We provide formal insights on the limitations of focusing solely on model-level fairness and highlight the need for alternative tools that account for heterogeneity in user preferences. To mitigate system-level disparities, we adapt closed-box optimization tools (e.g., BayesOpt) to jointly optimize utility and equity. We empirically demonstrate the effectiveness of our proposed framework on synthetic and real datasets, underscoring the need for a framework that reflects the design of modern, industrial AI systems.

Keywords: AI System, System-level Fairness, Fairness in Recommendation Systems

1. Introduction

The prevailing focus in algorithmic fairness is on bias measurement and mitigation for individual prediction models as the unit of analysis (Barocas et al., 2023; Mehrabi et al., 2022; Caton and Haas, 2020; Wan et al., 2023). However, industrial applications of ML rarely train and serve a single model in isolation: individual models are components of a broader AI system. The recent EU AI act defining its scope of an "AI system" as "*a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.*" This system-level perspective raises several questions for responsible development and deployment of recommendation systems. Are fairness notions for individual models sufficient to provide system-level equity?

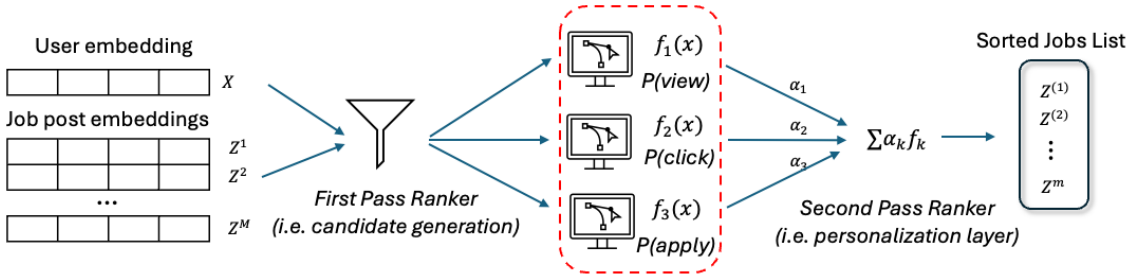

Figure 1: **AI Recommendation System Serving Pipeline.** Recommendations for feeds, ads, and social networking are generated from a multi-step process involving multiple ML models. An upstream process first fetches potentially relevant items via a candidate retrieval model, often called a "first pass ranker" (FPR), reducing the item set from millions to hundreds/thousands using scalable methods like approximate nearest neighbors (Liu et al., 2004). Then, each model f_k scores the items independently on based on probability of {View, Click, Apply}—this stage of ML models (red) are the overwhelming focus of fairness literature and audits. Finally, to surface the most relevant items, a "second pass ranker" (SPR) combines these individual models through a weighted sum $\sum_k \alpha_k f_k$. While the SPR formulation is simple, its intuitiveness has led to its ubiquity, with the world's largest platforms such as Meta (Vorotilov and Shugaepov, 2023), LinkedIn (Ouyang et al., 2022), Microsoft (Arunmozhi et al., 2021), X/Twitter (Twitter/X, 2023), Snapchat (Snapchat, 2022), Pinterest (Pinterest, 2020), and Spotify (Lamere, 2021) stating or suggesting that they use a variant of this overarching system.

We study AI system-level fairness in industrial recommendation systems, which are composed of multiple layers of ML models. In Figure 1, we illustrate a schema utilized by many of the largest tech companies in the world. Using a job recommendation system as a running example, the ultimate goal is hiring, as measured by the "confirmed hire rate" (see the left side of Figure 2). However, this system-level objective of "confirmed hire" is notably misaligned with the short-term predictions of individual models (e.g., the per-job clicks, views, applies). In turn, the fairness properties of those individual models may also be inadequate in addressing fairness when they are combined at the system level. AI systems frequently integrate additional serving rules such as priors on user preferences, known importance of different characteristics, prioritizing freshness of items.

We propose a system-level fairness framework for industrial recommendation engines that analyzes the entire pipeline, from retrieval to serving, rather than focusing on individual model fairness. This unified approach enables tracking bias propagation across components and its impact on utility disparities between user groups, particularly under preference heterogeneity ("distribution shifts"). We focus on the ultimate intent of the system, and using job recommendations as context, we develop a framework for measuring fairness through hiring outcome disparities.

Multi-objective black-box optimization for system-level fairness In most settings, the performance of individual ML models can be optimized "off-line" based on previously collected user data using standard tools like cross-validation. On the other hand, the FPR and SPR stages (Figure 1) require "online" experimentation (A/B testing) to collect user feedback on the quality of weights/parameters α. In the latter, weights are typically tuned through manual trial & error or black-box optimization methods such as Bayesian Optimization (BO) (Frazier, 2018). BO allows optimizing over noisy user feedback without derivative information, and is commonly used in online platforms (e.g., Meta (Letham et al., 2017) and LinkedIn (Agarwal et al., 2018)).

We go beyond identifying the source of disparities in utility, and formulate the overall system design as a multi-objective optimization problem balancing welfare and fairness. As we detail later, since it is difficult to model system-level objectives using a particular functional form, we treat them as a black-box and model it using a flexible Gaussian process. We use a simple but effective variant of a Bayesian Optimization (BO) algorithm to simultaneously maximize social welfare minus the disparity between the utility across groups. While the algorithms we leverage are well known, our formulation demonstrates how familiar tools from black-box optimization can be utilized in a practical and meaningful way to improve system-level fairness.

To further contextualize this in our job-recommendation setting, suppose that the job prediction model separately predicts two characteristics, *pClick* (proability of clicking) and *pApply* (the probability of applying). Disparities in job preferences across demographic groups could manifest as one group preferring to see attractive sounding jobs (click-worthy jobs), while another group preferring jobs of better background fit (apply-worthy jobs). When we have one model for clicks and one for applies and serve recommendations as a weighted linear combination of these model outputs, the "population weights" are unknown and are typically chosen "globally" such that all users receive the same weights. Yet, the ideal weights from a user experience perspective largely depend on individual preference, which traditional fairness metrics fail to capture. While weights can be personalized to certain demographic groups, this introduces disparate treatment concerns (Lipton et al., 2019; Seiner, 2006) and thus it is common practice to select a single global weight α. As we demonstrate on the right side of Figure 2, these global design choices have a tendency to tailor to the preferences of majority groups when using off-the-shelf tools.

In Section 4, we formulate a black-box optimization problem over global weights that avoids starkly benefiting one group over others. While our specific fair BO methodology is straightforward, it underscores the critical role that adaptive experimentation methodologies like BO can play in achieving system-level fairness for industrial applications.

Related work on viewer-side fairness in compositional systems In this work, we focus on *viewer-side fairness*, which studies the disparity in utility a recommender system provides to users. In particular, this ignores the agency of the items being recommended: if items represent human agents and/or their products (e.g., creators and their content), a more holistic approach is required that goes beyond the scope of the current paper.

For fairness at an individual model-level, several authors focus on group-wise disparity of ranking performance metrics like AUC, NDCG, or F1, ensuring "minimum quality of service." Examples include audits at Twitter (Lum et al., 2022), Microsoft (Microsoft,

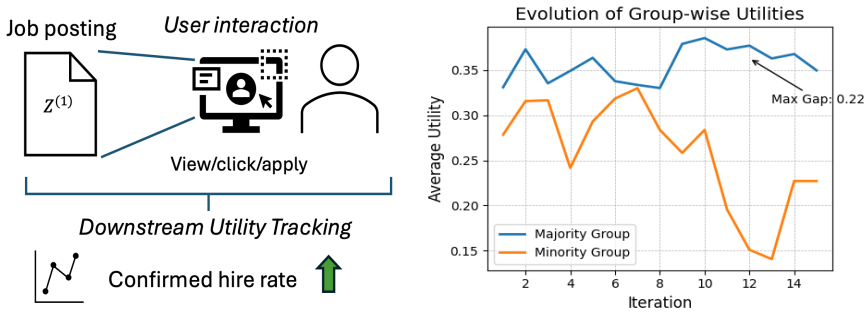

Figure 2: (Left) Proxy and metric tracking (Right) Issues of vanilla BayesOpt

2022), and LinkedIn (Quiñonero Candela et al., 2023), with comprehensive surveys by Li et al. (2023); Wang et al. (2023); Raj and Ekstrand (2022); Ekstrand et al. (2022). We root our analysis on the spirit of viewer-side fairness concepts in this work and focus on the group-wise disparity in overall viewer *utility*.

Fairness in compositional systems been studied from several perspectives; we summarize our key distinctions from existing works here and provide in-depth technical details and references in Appendix A. Our work differs from existing research on compositional fairness in two key aspects: (1) we propose a novel multi-label fairness framework grounded in practical utility modeling, departing from prior approaches that propose transformations from the multi-label setting to a single-label setting (Dwork and Ilvento, 2018; Wang et al., 2021a; Atwood et al., 2023). (2) While fairness in ML pipelines has previously been studied (Bower et al., 2017; Dwork et al., 2020; Blum et al., 2022; Khalili et al., 2021), we take this one step further and jointly analyze the filtration/selection layer in conjunction with the serving layer. We use this perspective to emphasize the need for fairness interventions at this stage by showing how it can represent a bottleneck in achieving fairness, even if everything else is fully fair.

Limitations Our main contribution is formulating and crystallizing the notion of fairness at the system-level, providing structural insights on how individual components can be optimized to achieve this goal. While formalizations provide practical value through concrete definitions and operational algorithms, our approach is inherently limited as system-level equity cannot be reduced to a single number. We highlight the need for multi-faceted approaches where i) stakeholders of the discrete components of the system collectively have incentives to prioritize equity at the system-level, and ii) a central organization coordinates system design with fairness as a central concern. Lastly, we make technical assumptions in Section 2 to focus on the single period interaction (rather than temporal) and focus on a specific variety of distribution shift. Handling these assumptions would introduce further granularity into the analysis that are tangential to our main message about system-level fairness and we consider these ideas as a topic for future study.

2. User Utility Under Compositional Recommender Systems

We first argue that fairness in AI systems should prioritize downstream user utility over individual model performance metrics. Although previous works have advocated for utility-centric fairness (Zehlike et al., 2022a,b; Singh and Joachims, 2019), these efforts do not

extend to compositional systems, where utility depends on multiple signals (i.e., model objectives). Crucially, reducing disparities in single-model fairness metrics (e.g., NDCG, AUC) does not necessarily improve system-level fairness, as proxy metrics often misalign with the system's ultimate utility goal. Instead of redefining fairness within individual models, we directly optimize fairness in downstream utility.

2.1. Framework for Compositional Utility

We formalize the notion of system-level utility we explore in this work. In a recommendation system with M items in its entire corpus (typically in the order of millions), let Z^j be the feature representations of each item $j = 1, \ldots, M$. When a user initiates a session with covariates X, the item/candidate retrieval step selects $m < M$ relevant items; let $I(X, Z^j) \in \{0, 1\}$ be the indicator for whether Z^j is selected. In practice, item Z^j may pass through rules-based filters or be chosen via a m-nearest neighbors algorithm.

For each item j and user (query) X, we have K different observable user outcomes Y_k for $k = 1, \ldots, K$ (K is usually ≤ 10, e.g., clicks, likes). These outcomes represent different aspects of item quality tracked by a domain expert (e.g., a product manager) who hypothesizes each Y_k is a proxy of the downstream utility metric U that they are tasked to improve (e.g., engagement, network growth). ML models $f_k : X, Z \to Y_k$ provide predictions of each outcome, which is combined with weights $\alpha_k > 0$ to recommend the best item $\hat{j}$ among the candidate set.

Definition 1 (Best Item for Serving) *Given a universe of M items, serving preferences $\alpha_k > 0$, model scores $f_k(X, Z^j)$, and retrieval function $I(X, Z^j)$, the recommended item $\hat{j}$ is given by*

$$\hat{j} := \mathit{argmax}_{1 \leq j \leq M} \left\{ \sum_{k=1}^{K} \alpha_k f_k(X, Z^j) I(X, Z^j) \right\} \tag{1}$$

Practitioners commonly use the weighted sum formulation (1) for interpretability and ease of adjustment. Even if the linear sum does not perfectly match user preferences, groups of users conceptually have an optimal set of weights within the linear model class. The universality of this model indicates that weighted sum is often viewed as a reasonable approximation of true user utility. Our framework generalizes to compositional functions (e.g., products, maxima), regression models, and cases involving negative utility (e.g., likelihood of abusive content) by introducing additional preference parameters. Although we focus on the top-1 item for clarity, our approach extends to top-k rankings, where position bias is an important consideration for future work.

To formalize user utility, we give the system designer the utmost benefit of the doubt and trust that their serving is "well-specified". In practical terms, we assume that the product owners have formulated the serving mechanism according to their best approximation for the abstract, downstream goal that they are aiming to optimize (e.g., number of job applicants). We shall see in the next section that even when we assume user utilities reflect the beliefs of domain experts, there can be large discrepancies between individual-model vs. system-level fairness.

Definition 2 (User Utility) *When item $\hat{j}$ is recommended, user X utility is given by the conditional expectation of the weighted sum under true outcomes and true preferences $\alpha^*(g)$ for group $g \in \mathcal{G}$*

$$U_g(X, I, f, \alpha, \alpha^*) = \mathbb{E}\left[\sum_{k=1}^{K} \alpha_k^\star(g) \cdot Y_k^{\hat{j}} \mid X\right] \tag{2}$$

Here, we implicitly assume X is rich enough that $\{Y_k^j\} \mid X$ is invariant across demographic groups $g \in \mathcal{G}$ and instead model group-level heterogeneities through the true preference vector $\alpha^*(g)$. Although users typically engage with the system multiple times, we treat them as i.i.d. in this work. We highlight this as a major limitation of our work as realistically, even a single bad experience can turn a user away from the platform.

Using this formalization, we can identify assumptions for when serving recommendations with $\alpha^*(g)$ is a maximizer of user utility. The conditions we require are that the scoring models f_k are fair in the sense that they are calibrated across the entire feature space. Calibration is an extensively studied topic in fairness literature (Hebert-Johnson et al., 2018; Błasiok et al., 2023; Liu et al., 2019; DiCiccio et al., 2023); though typically defined in case of candidate-side fairness, the same takeaways hold for the viewer-side case that we analyze.

Lemma 3 *Suppose individual models $f_k(X, Z^j)$ are calibrated with respect to their intended label Y_k across the entire feature space: $\mathbb{E}\left[Y_k^j \mid f_1(X, Z^j), \ldots, f_K(X, Z^j)\right] = f_k(X, Z^j)$ a.s. and that true and serving preferences are positive $\{\alpha^*, \alpha\} > 0$. If item j has a nonzero probability of being retrieved $\mathbb{P}\left(I(X, Z^j) = 1\right) > 0$ whenever $\mathbb{P}\left(f_k(X, Z^j) = \cdot \mid X, Z^j\right) > 0$ a.s., then setting $\alpha = c \cdot \alpha^*$ for $c > 0$ is a maximizer of utility.*

The lemma shows the validity of our framework in that learning true preferences indeed optimizes utility. We also qualitatively consider the role and impact of *unobservable* outcomes (e.g. sentimentality) in Appendix E.

3. Structural Insights

We now provide structural analyses of the utility gap between groups to answer the following question: *when is individual model fairness (in)sufficient for system-level fairness?* We articulate the different sources of utility-based inequity, and show the causes for utility gaps can be different compared to the standard causes of individual-model fairness. In particular, users who have the same features X look identical to the platform, but they may exhibit heterogeneous preferences (and therefore utilities) across groups. For example, different demographic groups may not respond identically to the same job posting despite similar backgrounds.

We consider two groups $G = 0$ and $G = 1$ to illustrate, and use $E_g[\cdot]$ to denote the expectation with respect to $\mathbb{P}(X = \cdot | G = g)$. Focusing on the user's short-term experience after a single interaction with the recommendation system, we analyze the utility gap

$$\mathbb{E}_1\left[U_1(X, I, f, \alpha, \alpha^*)\right] - \mathbb{E}_0\left[U_0(X, I, f, \alpha, \alpha^*)\right].$$

We assume the product *does not* personalize models specifically to demographic groups at any stage due to disparate treatment concerns—treating users differently based on their demographics. Without loss of generality, we denote group 0 to be the disadvantaged group.

Our goal is to perform an apples-to-apples comparison between demographic groups by "controlling" for the effects due to the user features X. That is, we wish to compare the gap utility $U_1(X, I, f, \alpha, \alpha^*) - U_0(X, I, f, \alpha, \alpha^*)$ for users that only differ in their group memberships. However, such a comparison is only possible over users co-observed in both groups. Thus, we define a notion of a "shared space" between $\mathbb{P}(X|G = 1)$ and $\mathbb{P}(X|G = 0)$

$$S_X(x) \propto (p(x|G = 1) + p(x|G = 0))^{-1} p(x|G = 1)p(x|G = 0), \tag{3}$$

so that S_X is small whenever either $p(x|G = 1)$ or $p(x|G = 0)$ is small and large when both quantities are large. Intuitively, taking expectations over S_X means we are paying attention to the feature space where both groups are present (e.g. industries X_1 where both demographics $G = 0, 1$ are represented). This mirrors Cai et al. (2023)'s distribution shift decomposition approach.

We expand the difference in expected utility as follows and provide some high level intuition:

$$\begin{aligned}
&\mathbb{E}_1 \left[U_1(X, I, f, \alpha, \alpha^*)\right] - \mathbb{E}_0 \left[U_0(X, I, f, \alpha, \alpha^*)\right] \\
&\quad = \mathbb{E}_1 \left[U_1(X, I, f, \alpha, \alpha^*)\right] - \mathbb{E}_{S_X}[U_1(X, I, f, \alpha, \alpha^*)] && (4) \\
&\qquad + \mathbb{E}_{S_X} \left[U_1(X, I, f, \alpha, \alpha^*) - U_0(X, I, f, \alpha, \alpha^*)\right] && (5) \\
&\qquad + \mathbb{E}_{S_X}[U_0(X, I, f, \alpha, \alpha^*)] - \mathbb{E}_0 \left[U_0(X, I, f, \alpha, \alpha^*)\right] && (6)
\end{aligned}$$

Term (4) relates to the utility change from the feature distribution of group 1 ($P(X|G = 1)$) to the shared distribution (3). It is large when the utility gap can be attributed to user features X often seen in group 1 but not in group 0. For instance, this may imply that the AI system provides better recommendations in male-dominated industries such as construction.

Term (5) compares the utility gap between groups over S_X. This term is large when the AI system favors the preferences α^* of the majority group 1. Finally, Term (6) is large if the AI system works better for users who are common in both groups compared to those only in group 0.

3.1. Impact of Preference Misspecification

We now take a closer look at Term (5). Further decomposition of these terms in Theorems 4, 6 to come unveil the impact and limitations of individual model fairness. We show that misspecification of user preferences in the serving model and disparities in the quality of the embedding model can also be significant drivers of utility gaps.

We analyze the "apples-to-apples" comparison in Term (5), which provides the most intuitive notion of "unfairness." This implies that the quality of recommendations is unequal even when two users from different groups share the same covariates (e.g. in interest and past behavior). We show this gap can occur at the system-level *even if the individual models are fair*. Below, the residuals $|Y_k^{\hat{j}} - f_k(X, Z^{\hat{j}})|$ measure model performance of f_k, where as $|\alpha_k - \alpha_k^*(0)|$ and $|\alpha_k - \alpha_k^*(1)|$ denote the estimation error in preferences of the two groups.

Theorem 4 (Utility Gap Bound From Preference Misspecification) *If recommendations are served using one set of α's for both groups, the expected utility gap (Term 5) is upper bounded by*

$$\sum_k \mathbb{E}_{S_X}\Big[|\alpha_k^*(1) - \alpha_k^*(0)| \cdot |Y_k^{\hat{j}} - f_k(X, Z^{\hat{j}})| \\ + |\alpha_k - \alpha_k^*(0)| \cdot f_k(X, Z^{\hat{j}}) + |\alpha_k - \alpha_k^*(1)| \cdot f_k(X, Z^{\hat{j}})\Big].$$

Typical fairness interventions aim to reduce the first term by ensuring each individual model performs well (Hebert-Johnson et al., 2018; Błasiok et al., 2023; Gopalan et al., 2021). Our bound affirms individual model fairness (specifically calibration) is a crucial requirement in AI system fairness, but also that disparities in model performance are amplified by differences in preferences across groups $\alpha_k^*(1) - \alpha_k^*(0)$, which are terms *not* reducible by the modeler. On the other hand, the second and third terms represent the utility disparity caused by misspecification of preferences. The utility disparity due to heterogeneous preferences has been been largely left out in the fairness literature; even prior works on compositional fairness treat all models/labels as equally important (Appendix A). We propose a concrete algorithmic approach in Section 4 to address this problem.

3.2. Downstream Impact of Upstream Candidate Selection Models

We now focus on Term (4), and move *upstream* of the scoring models to assess the impact of candidate retrieval model $I(X, Z)$ as a driver of utility gap. In practice, the candidate retrieval model is separate from the ML models and SPR serving layer and may even be managed by a separate engineering team. Retrieval models are often designed to optimize offline retrieval metrics such as recall over a single outcome (e.g., Click). Since engineering considerations such as latency, memory, and performance drive retrieval model design (Pinterest, 2021; Uber, 2023; Snapchat, 2023), fairness considerations are generally underappreciated.

While understanding user preferences over multiple labels is key to maximizing utility, candidate retrieval evaluations do not typically consider multiplicity of labels: recall over a single label may not align with the important label from the user's perspective. We address this by proposing a retrieval quality metric that we find more suitable for the compositional model system.

Definition 5 (Candidate Retrieval Model Quality) *A γ-good item j satisfies $\mathbb{E}\left[\sum_{k=1}^K Y_k^j\right] \geq \gamma$. The quality of a candidate retrieval model $I(X, Z)$ selecting m items is the expected highest γ-good item it can retrieve from the candidate pool. Formally, recalling $I(X, Z^j)$ is the indicator for whether item Z^j is retrieved for user X, the quality metric for the candidate retrieval model is*

$$Q_m(I(X, Z), Y) = \mathbb{E}\left[\max_{j \in \{1,\ldots,m\}} \left(I(X, Z_j) \cdot \sum_{k=1}^K Y_k^j\right)\right]$$

Notably, this definition differs from the standard definition of recall (true positives over total positives). This distinction is crucial for two reasons. First, this definition now addresses

downstream utility by spanning multiple labels that the business ultimately knows are relevant for user preferences. Second, we focus on the maximum because the downstream model and SPR layer are designed for surfacing the best item from the retrieved candidates.

With this definition in hand, we relate disparity in retrieval quality to that of utilities. We show that even if fully optimize utility with respect to everything *downstream* of the candidate retrieval model $I(X, Z)$, namely the scoring models f_k and serving coefficients α_k—the utility gap is still bottlenecked by the quality of the candidate retrievals.

Theorem 6 (Utility Gap Bound From Retrieval Performance Degradation) *Assume away biases from other sources so that $\alpha = \alpha^*$ and individual models are calibrated as in Lemma 3. For all users, let there be m^+ γ-good items in our item corpus of size M, and consider a ϵ-gap in retrieval quality $\mathbb{E}_1 [Q_m(I(X, Z), Y)] - \mathbb{E}_{S_X} [Q_m(I(X, Z), Y)] > \epsilon$. Then, Term (4) is at least*

$$\sup_{f,\alpha} \mathbb{E}_1 [U_1(X, I, f, \alpha, \alpha^*)] - \sup_{f,\alpha} \mathbb{E}_{S_X} [U_1(X, I, f, \alpha, \alpha^*)] \geq \epsilon \cdot \min_k \alpha_k^*(1).$$

In summary, our analysis emphasizes that solely assessing the fairness of candidate retrieval models in an offline setting (e.g., just measuring AUC/precision-recall) can mask their true impact on downstream utility, which is dictated by online metrics. Disparities introduced at the retrieval stage may persist even if subsequent models and serving layers are perfectly calibrated. This underscores the need for stakeholder alignment on system-wide objectives rather than isolated model-level metrics. More importantly, purely offline evaluations of retrieval models do not paint the full picture of downstream usefulness, making adaptive experimentation indispensable. By iteratively incorporating online feedback and aligning objectives across all parts of the recommendation pipeline, practitioners can more effectively optimize both utility and fairness in multi-stage AI systems.

4. System-Level Fairness Via Bayesian Optimization

We now shift from identifying fairness gaps to mitigating them. As discussed in Section 3.1, a key contributor to utility disparities is the overrepresentation of the majority group in determining preference weights α. We frame the selection of α as a derivative-free black-box optimization problem and propose an inequality-aware Bayesian Optimization (BO) approach. More broadly, we highlight that adaptive experimentation methods like BO hold great potential for system-level fairness.

4.1. Selecting an Inequality Metric

Prior work on fairness in BO has focused on standard fairness criteria like demographic parity or equalized odds (Perrone et al., 2021; Weerts et al., 2024; Candelieri et al., 2022; Sikdar et al., 2022). Instead, we advocate for utility-based fairness, moving beyond traditional definitions. While directly using the average utility gap across groups is intuitive, it lacks scale invariance, requires designating a disadvantaged group, and does not generalize well to multiple groups.

To address these issues, we adopt Deviation from Equal Representation (DER), introduced by Friedberg et al. (2022) to measure disparate utilities in experiments. For k groups

with non-negative downstream outcomes $\mu_1, \cdots, \mu_k$ (representing average number of sessions/confirmed hires per group), the DER is defined [1] as the following metric, which is scale invariant and naturally extends to multiple groups.

$$D(\mu_1, \cdots, \mu_k) = 1 - \frac{k}{k-1} \sum_k (\frac{\mu_k}{\sum_k \mu_k} - \frac{1}{k})^2 \tag{7}$$

When all means are equal, $D(\mu_1, \mu_2) = 0$ and otherwise the statistic grows larger when the values grow more disparate. Friedberg et al. (2022) originally used this to detect unintended biases in experiments; we integrate DER into BO to ensure that utility gains are not disproportionately concentrated in the majority group.

4.2. Incorporating Deviation from Equal Representation Into Bayesian Optimization for System-Level Fairness

A common strategy to considering fairness in BO is to use a constrained version of the expected improvement (EI) methodology by Gardner et al. (2014); Perrone et al. (2021), where performance and fairness criteria are modeled as separate, independent Gaussian Processes (GPs) with one GP representing the objective and the other as the constraint. This method is interpretable and computationally efficient, but applying it directly to DER poses two challenges. First, DER is not independent of the overall utility objective, invalidating the assumptions behind constrained EI. Second, setting an appropriate fairness threshold is nontrivial, as business objectives often seek to improve both DER and overall utility simultaneously rather than enforcing a hard constraint. We provide a more in-depth discussion of EI and its limitations in B.

Instead of constrained EI, we frame the problem as a multi-objective optimization and use *Expected Hyper-Volume Improvement* (EHVI) (Daulton et al., 2020; Yang et al., 2019). EHVI is the multi-objective analog of EI, optimizing improvements in Pareto frontier hypervolume rather than a single scalar objective. By optimizing for the global utility and DER simultaneously, we aim to find points that present the best possible tradeoffs. At each iteration, EHVI returns a set of preference weights that have the highest expectation of hypervolume improvement and we prioritize those with the highest expected DER. This allows us to find points that encourage fairness while also boosting global utility.

5. Experiments

5.1. Datasets

We now demonstrate our proposed multi-objective-optimization formulation of utility and DER as a simple but effective method for achieving system-level fairness. Due to the lack of public benchmarks with readily available multi-action outcomes to credibly represent an industrial system, we restrict our analysis to four datasets `MovieLens`(Harper and Konstan, 2015), `ModCloth`(Wan et al., 2019), `Electronics`(Wan et al., 2019), and one synthetic example. These datasets represent varying levels of difficulty in terms of model performance,

1. Although DER is originally defined as 1 minus the quantity shown, we renormalize it to have the interpretation that "higher is more equitable/better" to better align with the canonical view of Bayesian optimization.

utility optimization, and DER optimization. Details are provided in Appendix D. Importantly, we *start* with models that are groupwise fair, which is typically the end-goal of fairness studies. We then simulate group-wise preference disparities and use BOTorch (Balandat et al., 2020) for iterative Bayesian optimization to mimic online recommendation systems. Each iteration samples users, retrieves m items per user, and computes utility based on true labels and preferences as defined in Definition 2.

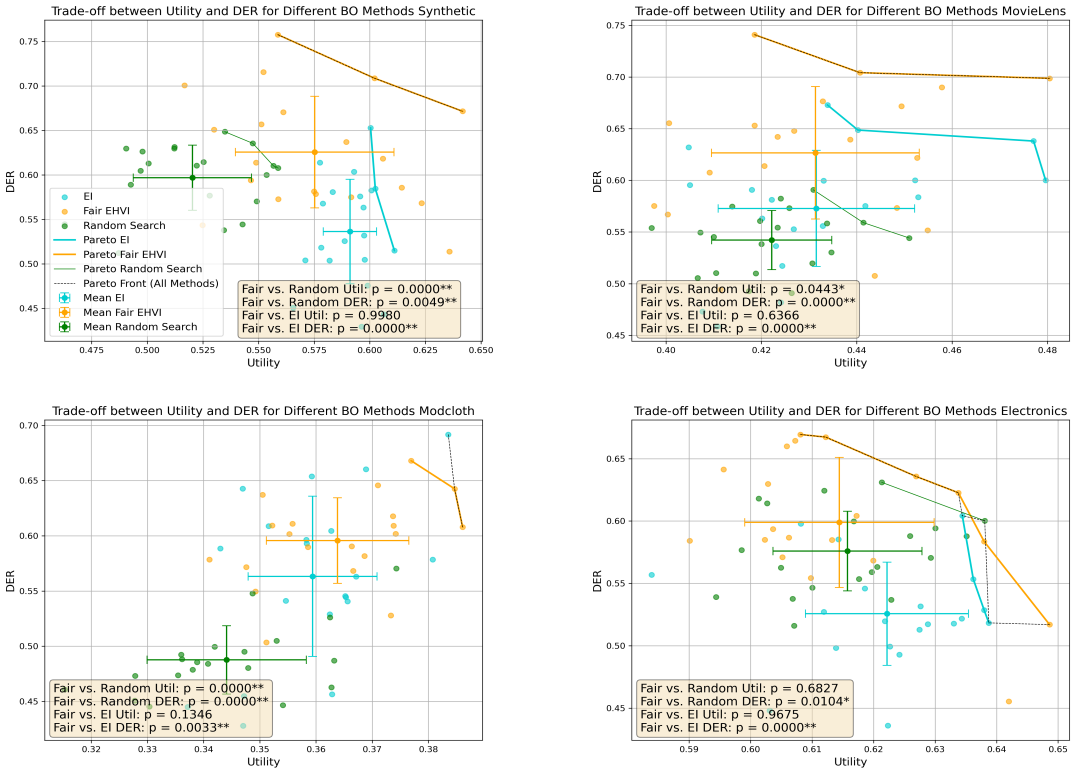

Figure 4: Pareto frontiers for the four tested datasets

5.1.1. Results and Comparison

For each dataset and method, we aggregate the utility and DER over the 20 iterations for each of the 20 trials. We trace the Pareto frontier for each method across the trials, and then plot the average utility and DER to represent the average-case performance with 1 stdev error bars. For statistical significance, we report the p-value of the Wilcoxon signed rank test comparing our Fair EHVI against the baseline (random search) and pure utility optimization (EI). We consider random search the baseline, as even simple BO processes like EI are non-trivial implement in industry and therefore may not be available. All experiments were run locally on a MacBook Pro with an M3 processor.

In Figure 3, we see that Fair EHVI overall beats both random search and EI in identify the best tradeoff between utility and DER. For a more detailed view of performance on each dataset, we turn to Figure 4. The plots demon-

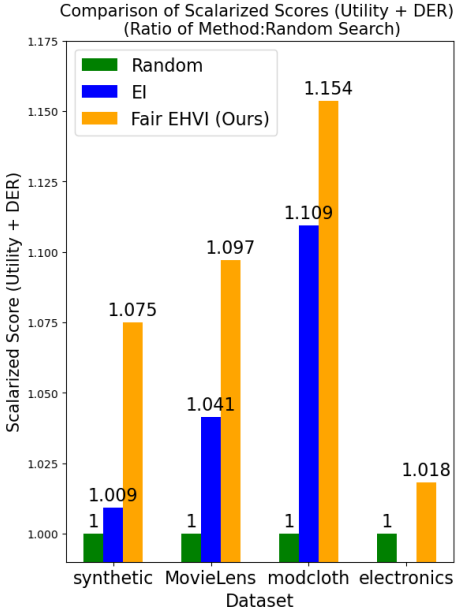

Figure 3: Comparison of methods on scalarized outcome

strate that Fair EHVI consistently yields the most points on the global pareto frontier of utility and DER. Fair EHVI also yields better DER than random search and EI by a statistically significant margin in all cases, even in the difficult Electronics dataset where random search is difficult to beat. Our methodology does suffer in that dataset where the average utility is slightly lower than random search, but this can potentially be overcome with more tuning of the search parameters (as our algorithm samples EHVI candidates in parallel and optimizes for DER). Overall, we reiterate that the significance of adaptive experimentation tools like BO in AI systems necessitates further research on using them to trade between fairness and business objectives, or even between fairness objectives (Hsu et al., 2024; Bell et al., 2023).

6. Conclusion and Future Work

In this paper, we have proposed a mechanism for shifting from model-centric to AI system-level fairness. We align fairness measurement with the system's objective of user utility optimization, spotlight the retrieval and serving layers as critical points of intervention, demonstrate the benefits and limitations of intervention at these layers, and propose a Bayesian optimization solution for bias mitigation at serving time. Our empirical results show improved utility distribution across heterogeneous user groups. As regulations like the EU AI Act emerge, understanding the framework for holistic AI system-level analysis becomes crucial for responsible AI practices. Even beyond the algorithmic-focused interventions that we have presented, we propose that a crucial first step is simply recognizing the different parts of an AI system - retrieval, models, serving layers, rerankers/business rules - and documenting their respective purposes and potential influences on the disparity in user utility.

For future work, one critical direction is understanding how the timescales of utilities affect fairness. For instance, while we have assumed that either all users are unique or that user sessions are i.i.d, this is generally not true. High utility in one session begets further usage and low utility begets strong drop-off. While these patterns can only be empirically estimated, folding them into our fairness framework adds the timescale dimension of how long unfairness is tolerable by users, further motivating the urgency of the problem. Additionally, as we have framed system-level fairness for the viewer-side in this paper, we encourage researchers to understand the analogous problem but for systems that rank candidates and the interaction between the two types of fairness.

Acknowledgments

We sincerely thank Sam Gong and Will Cai for their insightful feedback and in-depth conversations about system-level fairness. We would also like to thank Sakshi Jain and Heloise Logan for their support and Kinjal Basu for his improvements on the DER metric. Finally, we thank the anonymous reviewers for their helpful comments on notation improvements and for providing references to expand the discussion of related works.

References

Deepak Agarwal, Kinjal Basu, Souvik Ghosh, Ying Xuan, Yang Yang, and Liang Zhang. In *Online Parameter Selection for Web-based Ranking Problems*, KDD '18, page 23–32, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520.

Madhulekha Arunmozhi, Ian Ackerman, Manas Somaiya, and Prashant Saxena. Multistage feed ranking system with methodology providing scoring model optimization for scaling, Dec 2021.

James Atwood, Tina Tian, Ben Packer, Meghana Deodhar, Jilin Chen, Alex Beutel, Flavien Prost, and Ahmad Beirami. Towards a scalable solution for improving multi-group fairness in compositional classification. 2023. URL https://arxiv.org/pdf/2307.05728.pdf.

Maximilian Balandat, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. Botorch: A framework for efficient monte-carlo bayesian optimization, 2020.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and machine learning: Limitations and opportunities. 2023.

Andrew Bell, Lucius Bynum, Nazarii Drushchak, Tetiana Zakharchenko, Lucas Rosenblatt, and Julia Stoyanovich. The possibility of fairness: Revisiting the impossibility theorem in practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 400–422, 2023.

Avrim Blum, Kevin Stangl, and Ali Vakilian. Multi stage screening: Enforcing fairness and maximizing efficiency in a pre-existing pipeline. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1178–1193, 2022.

Amanda Bower, Sarah N. Kitchen, Laura Niss, Martin J. Strauss, Alexander Vargas, and Suresh Venkatasubramanian. Fair pipelines, 2017. URL https://arxiv.org/abs/1707.00391.

Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, Adam Tauman Kalai, and Preetum Nakkiran. Loss minimization yields multicalibration for large neural networks, 2023.

Tiffany Tianhui Cai, Hongseok Namkoong, and Steve Yadlowsky. Diagnosing model performance under distribution shift, 2023.

William Cai, Ro Encarnacion, Bobbie Chern, Sam Corbett-Davies, Miranda Bogen, Stevie Bergman, and Sharad Goel. Adaptive sampling strategies to construct equitable training datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1467–1478, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533203. URL https://doi.org/10.1145/3531146.3533203.

Antonio Candelieri, Andrea Ponti, and Francesco Archetti. Fair and green hyperparameter optimization via multi-objective and multiple information source bayesian optimization, 2022.

Simon Caton and Christian Haas. Fairness in machine learning: A survey, 2020.

Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. Differentiable expected hypervolume improvement for parallel multi-objective bayesian optimization, 2020.

Cyrus DiCiccio, Brian Hsu, Yinyin Yu, Preetam Nandy, and Kinjal Basu. Detection and mitigation of algorithmic bias via predictive parity. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1801–1816, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924.

Cynthia Dwork and Christina Ilvento. Fairness under composition. *ArXiv*, abs/1806.06122, 2018. URL https://api.semanticscholar.org/CorpusID:49303187.

Cynthia Dwork, Christina Ilvento, and Meena Jagadeesan. Individual fairness in pipelines. *arXiv preprint arXiv:2004.05167*, 2020.

Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. Fairness in information access systems. *Foundations and Trends® in Information Retrieval*, 16(1–2):1–177, 2022. ISSN 1554-0677. doi: 10.1561/1500000079. URL http://dx.doi.org/10.1561/1500000079.

Peter I. Frazier. A tutorial on bayesian optimization, 2018.

Rina Friedberg, Stuart Ambler, and Guillaume Saint-Jacques. Representation-aware experimentation: Group inequality analysis for a/b testing and alerting. *arXiv preprint arXiv:2204.12011*, 2022.

Jacob R Gardner, Matt J Kusner, Zhixiang Eddie Xu, Kilian Q Weinberger, and John P Cunningham. Bayesian optimization with inequality constraints. In *ICML*, volume 2014, pages 937–945, 2014.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors, 2021.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1939–1948. PMLR, 10–15 Jul 2018.

Brian Hsu, Rahul Mazumder, Preetam Nandy, and Kinjal Basu. Pushing the limits of fairness impossibility: who's the fairest of them all? In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713871088.

Mohammad Mahdi Khalili, Xueru Zhang, and Mahed Abroshan. Fair sequential selection using supervised learning models. *Advances in Neural Information Processing Systems*, 34:28144–28155, 2021.

Jen Lamere. A look behind blend: The personalized playlist for you...and you, Dec 2021. URL https://engineering.atspotify.com/2021/12/a-look-behind-blend-the-personalized-playlist-for-youand-you/.

Benjamin Letham, Brian Karrer, Guilherme Ottoni, and Eytan Bakshy. Constrained bayesian optimization with noisy experiments. *Bayesian Analysis*, 14, 06 2017. doi: 10.1214/18-BA1110.

Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. Fairness in recommendation: Foundations, methods and applications, 2023.

Zachary C. Lipton, Alexandra Chouldechova, and Julian McAuley. Does mitigating ml's impact disparity require treatment disparity?, 2019.

Lydia T. Liu, Max Simchowitz, and Moritz Hardt. The implicit fairness criterion of unconstrained learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4051–4060. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/liu19f.html.

Ting Liu, Andrew Moore, Ke Yang, and Alexander Gray. An investigation of practical approximate nearest neighbor algorithms. *Advances in neural information processing systems*, 17, 2004.

Kristian Lum, Yunfeng Zhang, and Amanda Bower. De-biasing "bias" measurement. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22. ACM, June 2022. doi: 10.1145/3531146.3533105. URL http://dx.doi.org/10.1145/3531146.3533105.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning, 2022.

Microsoft, Jun 2022. URL https://www.microsoft.com/en-us/ai/principles-and-approach.

Yunbo Ouyang, Viral Gupta, Kinjal Basu, Cyrus Diccicio, Brendan Gavin, and Lin Guo. Using bayesian optimization for balancing metrics in recommendation systems, Feb 2022. URL https://www.linkedin.com/blog/engineering/recommendations/using-bayesian-optimization-for-balancing-metrics-in-recommendat.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Valerio Perrone, Michele Donini, Muhammad Bilal Zafar, Robin Schmucker, Krishnaram Kenthapadi, and Cédric Archambeau. Fair bayesian optimization. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 854–863, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462629. URL https://doi.org/10.1145/3461702.3462629.

Pinterest. Improving the quality of recommended pins with lightweight ranking, Sep 2020. URL https://medium.com/pinterest-engineering/improving-the-quality-of-recommended-pins-with-lightweight-ranking-8ff5477b20e3.

Pinterest. Pinterest home feed unified lightweight scoring: A two-tower approach, Sep 2021. URL https://medium.com/pinterest-engineering/pinterest-home-feed-unified-lightweight-scoring-a-two-tower-approach-b3143ac70b55.

Joaquin Quiñonero Candela, Yuwen Wu, Brian Hsu, Sakshi Jain, Jennifer Ramos, Jon Adams, Robert Hallman, and Kinjal Basu. Disentangling and operationalizing ai fairness at linkedin. FAccT '23, page 1213–1228, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594075. URL https://doi.org/10.1145/3593013.3594075.

Amifa Raj and Michael D. Ekstrand. Measuring fairness in ranked results: An analytical and empirical comparison. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 726–736, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3532018. URL https://doi.org/10.1145/3477495.3532018.

Joseph Seiner. Disentangling disparate impact and disparate treatment: Adapting the canadian approach. *Yale L. & Pol'y Rev.*, 25:95, 2006.

Sandipan Sikdar, Florian Lemmerich, and Markus Strohmaier. Getfair: Generalized fairness tuning of classification models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 289–299, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533094. URL https://doi.org/10.1145/3531146.3533094.

Ashudeep Singh and Thorsten Joachims. Policy learning for fairness in ranking. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/9e82757e9a1c12cb710ad680db11f6f1-Paper.pdf.

Snapchat. Machine learning for snapchat ad ranking, Feb 2022. URL https://eng.snap.com/machine-learning-snap-ad-ranking.

Snapchat. Embedding-based retrieval with two-tower models in spotlight, Jun 2023. URL https://eng.snap.com/embedding-based-retrieval.

Twitter/X, Mar 2023. URL https://blog.x.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm.

Uber. Innovative recommendation applications using two tower embeddings at uber — uber blog, Jul 2023. URL https://www.uber.com/en-CA/blog/innovative-recommendation-applications-using-two-tower-embeddings/.

Vladislav Vorotilov and Ilnur Shugaepov. Scaling the instagram explore recommendations system, Aug 2023. URL https://engineering.fb.com/2023/08/09/ml-applications/scaling-instagram-explore-recommendations-system/.

Mengting Wan, Jianmo Ni, Rishabh Misra, and Julian McAuley. Addressing marketing bias in product recommendations, 2019. URL https://arxiv.org/abs/1912.01799.

Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. In-processing modeling techniques for machine learning fairness: A survey. *ACM Trans. Knowl. Discov. Data*, 17(3), mar 2023. ISSN 1556-4681. doi: 10.1145/3551390. URL https://doi.org/10.1145/3551390.

Xuezhi Wang, Nithum Thain, Anu Sinha, Flavien Prost, Ed H Chi, Jilin Chen, and Alex Beutel. Practical compositional fairness: Understanding fairness in multi-component recommender systems. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 436–444, 2021a.

Xuezhi Wang, Nithum Thain, Anu Aradhana Sinha, Flavien Prost, Ed H. Chi, Jilin Chen, and Alex Beutel. Practical compositional fairness: Understanding fairness in multi-component recommender systems. In *WSDM 2021*, 2021b. URL https://arxiv.org/pdf/1911.01916.pdf.

Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A survey on the fairness of recommender systems. 41(3), feb 2023. ISSN 1046-8188. doi: 10.1145/3547333. URL https://doi.org/10.1145/3547333.

Hilde Weerts, Florian Pfisterer, Matthias Feurer, Katharina Eggensperger, Edward Bergman, Noor Awad, Joaquin Vanschoren, Mykola Pechenizkiy, Bernd Bischl, and Frank Hutter. Can fairness be automated? guidelines and opportunities for fairness-aware automl. *Journal of Artificial Intelligence Research*, 79:639–677, February 2024. ISSN 1076-9757. doi: 10.1613/jair.1.14747. URL http://dx.doi.org/10.1613/jair.1.14747.

Kaifeng Yang, Michael Emmerich, André Deutz, and Thomas Bäck. Multi-objective bayesian global optimization using expected hypervolume improvement gradient. *Swarm and Evolutionary Computation*, 44:945–956, 2019. ISSN 2210-6502. doi: https://doi.org/10.1016/j.swevo.2018.10.007. URL https://www.sciencedirect.com/science/article/pii/S2210650217307861.

Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking, part i: Score-based ranking. *ACM Comput. Surv.*, 55(6), dec 2022a. ISSN 0360-0300. doi: 10.1145/3533379. URL https://doi.org/10.1145/3533379.

Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking, part ii: Learning-to-rank and recommender systems. *ACM Comput. Surv.*, 55(6), dec 2022b. ISSN 0360-0300. doi: 10.1145/3533380. URL https://doi.org/10.1145/3533380.