Scoping Fairness Objectives and Identifying Fairness Metrics for Recommender Systems: The Practitioners' Perspective

Jessie J. Smith
jessie.smith-1@colorado.edu
University of Colorado, Boulder
USA

Lex Beattie
lex@spotify.com
Spotify
USA

Henriette Cramer
henriette@spotify.com
Spotify
USA

ABSTRACT

Measuring and assessing the impact and "fairness" of recommendation algorithms is central to responsible recommendation efforts. However, the complexity of fairness definitions and the proliferation of fairness metrics in research literature have led to a complex decision-making space. This environment makes it challenging for practitioners to operationalize and pick metrics that work within their unique context. This suggests that practitioners require more decision-making support, but it is not clear what type of support would be beneficial. We conducted a literature review of 24 papers to gather metrics introduced by the research community for measuring fairness in recommendation and ranking systems. We organized these metrics into a 'decision-tree style' support framework designed to help practitioners scope fairness objectives and identify fairness metrics relevant to their recommendation domain and application context. To explore the feasibility of this approach, we conducted 15 semi-structured interviews using this framework to assess which challenges practitioners may face when scoping fairness objectives and metrics for their system, and which further support may be needed beyond such tools.

CCS CONCEPTS

• **Human-centered computing**;

KEYWORDS

fairness, recommender systems, algorithmic bias, algorithmic audits

ACM Reference Format:
Jessie J. Smith, Lex Beattie, and Henriette Cramer. 2023. Scoping Fairness Objectives and Identifying Fairness Metrics for Recommender Systems: The Practitioners' Perspective. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3543507.3583204

1 INTRODUCTION

Recommendation and ranking systems have become extremely prolific on the web. These often personalized systems leverage algorithms to recommend content, items, or information that matches users' perceived preferences. However, previous work has highlighted how personalized systems might also lead to unintentional harm, such as degenerate feedback loops [23, 43], sexist stereotyping [21], or racial bias [1]. This realization has resulted in calls to action for industry to audit, understand, and mitigate the potential algorithmic impact of their recommendation systems. However, it is not always clear for practitioners exactly how to do this, nor easy to keep track of all related literature. One aspect of algorithmic impact measurement is the proliferation of so-called 'fairness' metrics to choose from. This paper introduces a decision framework aimed at helping practitioners scope and identify such metrics for their given context. We define the process of scoping fairness contexts and constraints, identifying proxies to measure fairness, and implementing those measurements in practice as **operationalizing**.

Scoping fairness in recommendations from a theoretical qualitative construct to a quantitative measurement is not an easy task [22, 36]. Many **fairness definitions** have been introduced and explored by the recommender system research community (e.g., [15, 18, 25, 28, 41]). Fairness definitions (or objectives, goals) consist of different restrictions, assumptions, and requirements that must be met in order for a machine learning (ML) model to be classified as "fair" under that definition. Note however that a model itself being classified as fair does not guarantee fairness of its wider context. This combination of requirements for a given fairness definition can be referred to as **fairness constraints**. Each fairness definition has associated constraints that might differ depending on the context of the system and its users, the goals of stakeholders around models, and aspects such as the structure of training or evaluation data. Thus, defining and refining what constraints need to be met for a given fairness context can be a complex process that involves experts from different disciplines [33, 34]. For example, a web-based system recommending jobs to job-seekers may have very different fairness considerations than one recommending books to readers. Both systems might implement a similar metric, but their context, constraints, and goals might differ substantially.

Jobin et al. [24] described the proliferation of AI ethics guidelines and found more than 80 documents containing principles or guidelines, pointing to the need for more practical guidance beyond principles. As the field has moved to operationalize principles into practice, a multitude of complex, contextual fairness metrics have been introduced. The wide variety of available metrics, coupled with the lack of accepted standards or shared practice knowledge [16, 44], leads to a challenging environment for practitioners to navigate. These challenges can also lead to a widening gap between the research community and practitioners concerning the availability of metrics versus the ability to put them into practice.

We specifically focus on the challenges that practitioners encounter when scoping qualitative fairness objectives and identifying metrics for evaluating their real-world recommender systems,

WWW '23, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04.
https://doi.org/10.1145/3543507.3583204

before implementation begins. We describe how challenges during these two steps might prevent practitioners from operationalization and propose future work to address these barriers. During this study, we focused on both recommendation and ranking algorithms; we use these words interchangeably throughout this paper.

We conducted a literature review of 24 papers on recommendation fairness to design a preliminary framework for scoping qualitative and quantitative fairness objectives with related fairness metrics. This framework consists of a **decision tree** to help practitioners scope a potential harm in fairness terms and identify related quantitative metrics. We then conducted semi-structured interviews with 15 ML or ML-adjacent practitioners working in a recommendation setting. We presented interview participants with our scoping framework and facilitated feedback while observing their use of the framework in the context of their recommendation systems. Leveraging our framework allowed us to observe how participants interacted with ideas, definitions, and methods commonly researched within this space. For the purpose of this paper, we present the framework as a case study to help us understand what on-the-ground challenges practitioners face when scoping fairness-related harms and identifying fairness metrics in practice and what support or guidance is most helpful for various decision-making roles within a large tech organization. Specifically, in this work, we explore the following research question:

> **What guidance do ML recommendation practitioners need when scoping fairness objectives and identifying related fairness metrics?**

Our main contributions in this work are threefold: (1) a literature review of recommendation and ranking fairness metrics; (2) a preliminary framework organizing said literature to help practitioners scope fairness objectives and identify fairness metrics; and (3) results from semi-structured interviews with practitioners.

Similar to previous observations on fairness in recommendation systems in practice [3], we found that there are no standard decision-making processes for practitioner teams to approach operationalizing fairness metrics in recommendation systems. We also learned that practitioners need more guidance when defining fairness objectives and scoping those objectives into their associated metrics. We discuss how to address these findings based on the current state of research on fairness in recommendation systems and suggestions from participants in our study.

2 BACKGROUND

In this work, we define **qualitative fairness objectives** as fairness goal statements, such as those surfacing from an algorithmic impact assessment of a machine learning system [34]. In contrast, **quantitative fairness objectives** reflect quantitative definitions of fairness found in the literature, which often are accompanied by objective-specific fairness metrics (such as demographic parity or top-k fairness) [53]. We note that "fairness" objectives and metrics are often more closely related to minimizing "impacts" or "harms" caused by a system, yet the term "fair" has been widely adopted in the literature; fairness metrics may not necessarily capture all algorithmic impact questions (e.g., fairness metrics may not address recommendations' potential direct and indirect cultural impact).

Thus, a starting point for measuring fairness in a machine learning system is to declare what type of unfair impact or *harm* to measure. Crawford [14] previously categorized harms in machine learning as two broad types: (1) harms of **allocation**; and (2) harms of **representation**. In recommender systems, harms of allocation might refer to a system's unfair distribution of attention or exposure of items, while harms of representation might refer to a system's unfair representation of reality for a given information space [18]. Harm, impact, and unfair treatment are all related in machine learning systems and, in some cases, can be empirically observed through direct or proxy fairness metrics.

After determining a potential harm to target, practitioners will need to "scope" an appropriate fairness proxy for their context. This is a crucial step due to fairness being a subjective and non-observable construct [22, 49]. This paper investigates whether and how a framework for mapping qualitative fairness objectives to quantitative measurements can help practitioners determine what types of fairness metrics are most relevant in a given context.

2.1 Fairness Metric Constraints

The ML research community has steadily introduced and explored new fairness definitions, and their associated metrics for responsible AI efforts [2, 38]. A large portion of this work on fairness definitions and metrics for machine learning has focused on the domain of (binary) classification and regression. However, measuring fairness in recommendation systems poses distinct fairness challenges that are not shared with classification and/or regression systems [15, 18, 25, 41]. This has resulted in the creation of a specific area of research focusing on fairness within a recommendation or ranking context. Here we outline some of the unique challenges for recommendation systems in the form of "constraints" (or restrictions) that must be chosen in order to appropriately align a fairness metric with a specific fairness context.

2.1.1 The Multistakeholder Constraint. Recommendation system fairness is commonly referred to as **multistakeholder** fairness due to the goal of satisfying the fairness needs of multiple groups of stakeholders [9]. The two most common stakeholder groups are **providers** (those who provide or create content to be recommended) and **consumers** (those who interact with or consume the recommendations) [9]. Consumer-side or provider-side fairness can sometimes conflict with one another [9].

2.1.2 The Group vs. Individual Constraint. Similar to measuring fairness in other domains of machine learning, evaluating fairness in recommendation systems could require differentiating between measuring group and individual fairness [17]. Recent work has highlighted that group and individual fairness both stem from similar principles, but differ when measuring them in practice [6]. Group fairness measures if different groups of providers or consumers acquire similar recommendation outcomes, while individual fairness requires that similar individuals are treated similarly. In recommendation and ranking, different metrics can measure group versus individual fairness within each stakeholder category; some individual metrics can also be adapted to measure different kinds of individual or group fairness [18].

2.1.3 The System Component Constraint. Recommendation systems often consist of multiple components which leverage unique algorithms and sub-systems to create final recommendations for consumption. System components could include generating and retrieving pools of content, filtering said pools by ranking for high-priority goals, and finally re-ranking the final lists of content for consumption. This composite of system components results in a need to select at which step(s) to measure, while also understanding how biases may change across the system [32, 50, 51, 56]. Fairness could be measured over the entire population of recommended items (e.g., against a target distribution) or as a ranking problem (e.g., are the top-k rankings satisfying our fairness goals?). These types of choices can also have consequences for results and potential processing necessary [8]. This constraint compounds the complexity of choosing fairness metrics for practitioners.

2.1.4 The Fairness Objective Constraint. Previous work has attempted to categorize fairness metrics based on the fairness proxies being measured. Verma et al. [53] classified RecSys fairness metrics as *accuracy* based, *error* based, and *causal* based. This paper classified the majority of said metrics as accuracy based, where the metrics *"either state the condition for user satisfaction or provide a measure of deviation from the ideal ranking."* Verma et al. [53] describes pairwise fairness metrics, introduced by Kuhlman et al. [26], as error based metrics due to their measurement of false negatives and false positives against assumed ground-truth rankings. Causal based metrics require that item ranking is not related to group membership. More recently, Ekstrand et al. [18] published an in-depth review of fairness in recommendation systems, introducing more nuanced ways to classify metrics beyond the three categories originally suggested by Verma et al. [53]. Notably, Ekstrand et al. [18] categorized pairwise fairness metrics with accuracy metrics, alleviating the potential confusion between distinguishing when a metric measures error versus accuracy. The differences between these two publications reflect how fairness literature may change over time, making it difficult for practitioners to stay up-to-date and navigate this complex research space.

2.2 Challenges in Industry

Beyond the practical challenges practitioners face when selecting appropriate fairness constraints, barriers exist that can make it challenging for practitioners to operationalize algorithmic responsibility in industry web technologies. Rakova et al. [42] previously noted that apart from practical challenges in implementation, *"[responsible] AI initiatives also require operationalization within — or around — existing corporate structures and organizational change."* Integrating fairness work into existing corporate or business structures can sometimes be a difficult task, especially considering trade-offs that may occur [9]. Perceived alignment with existing business goals and resolving perceived tensions is vital [42]. Tensions also appear in advice within the literature. For example, Holstein et al. [20] discuss the gap between academic research and practical implementation. Specifically, this research highlighted that domain-specific, scalable standards would be greatly beneficial, as also discussed by [19, 29]. However, some research has objected to scaling, standardizing, or automating ML fairness in practice, due to the context-dependent nature of fairness work [30, 54].

Researchers and practitioners have turned towards tooling to combat some of these challenges. **Tooling** in the context of this work includes open-source libraries that provide code to implement fairness metrics, as well as protocols, questionnaires, and worksheets to help educate and support decision-making in an algorithmic responsibility setting [4, 7, 12, 34, 37, 44, 55]. However, despite all of these available tools, practitioners still report a disconnect between these tooling efforts and what they actually need in practice [27, 31, 33, 44, 52].

One major critique of ML fairness tools is that they do not provide the necessary guidance for practitioners to *begin* exploring their fairness goals [31]. Open-source libraries that provide metrics for practitioners are of little use if they do not know how to define the impact they want to measure, or scope their measurement context. Additionally, most ML practitioners are not trained in disciplines like ethics or philosophy [47], which creates another barrier to entry for this complex decision-making space. Scoping fairness goals and constraints might seem very daunting without institutional or academic knowledge of disparate impact or fairness metrics. In response to these challenges, Saleiro et al. [46] created a decision tree to help practitioners select an ML fairness metric. However, this decision tree assumes that the practitioner has prior knowledge of policy and ethics jargon, with some branches in the tree asking questions like, *"are your interventions punitive or assistive."* Additionally, this decision tree was designed for the context of binary classification, not ranking or recommendations. In the context of recommendation systems, fairness metrics and considerations are vastly different from a binary classification setting, especially since outcomes are not necessarily binary nor measurably favorable. There is rarely a "ground-truth" to compare the final recommendation lists against beyond assuming user engagement as a positive prediction [5]. To combat some of these challenges, we aimed to iteratively create an easier-to-understand, recommender-specific fairness scoping framework. We explore how practitioners interacted with this framework and how their feedback informed its design in Section 4, and opportunities for improvement.

3 METHODS

To capture the complex challenges practitioners face when scoping and identifying fairness metrics, we iteratively designed a decision-making framework based on a literature review and feedback from our participants. Our framework is a decision tree designed to specifically scope quantifiable harms and corresponding metrics from a pre-defined, conceptual, potential harm of the system. The **Decision Tree** helps practitioners decide between various fairness constraints to scope which quantitative fairness context and metric category is most appropriate for their goals.

These artifacts enabled us to guide practitioner interviews and categorize challenges they were encountering, and which part of the framework caused confusion or spurred feedback. We created this decision tree to mirror past frameworks introduced by Aequitas and Fairlearn, which helped scope quantitative fairness contexts and identify fairness metrics for binary classification and regression [7, 46]. To the best of our knowledge, we were unable to find versions of these types of tools for recommendation or ranking systems.

We observed challenges that practitioners may encounter by conducting guided interviews where practitioners mapped a potential fairness-related harm to avoid, to a group of potential metrics for their specific context. An overview of our interview process and how it incorporated the use of these artifacts is shown in Figure 1.

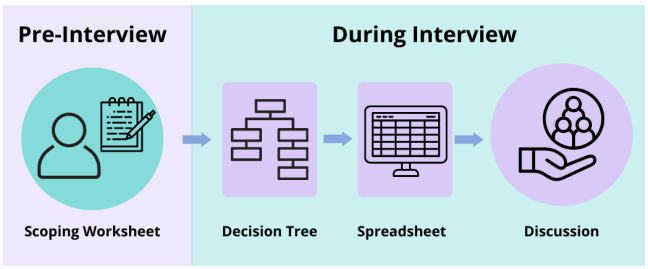


Figure 1: Interview process and artifacts used.

To glean more nuanced observations from interviews, we iterated on the framework to account for participant feedback. Our iterative design process was inspired by previous research on co-designing ML fairness standards (e.g., [34]); upon iterations, we addressed practical challenges that participants had explicitly mentioned or that we had implicitly observed. Our framework iterations allowed us to observe less obvious nuances in how practitioners operationalize fairness in later interviews in the study.

3.1 Designing the Original Prototype

Before conducting interviews, we created a low-fidelity prototype of our tools. The purpose of these prototypes was to help us uncover the practical challenges that practitioners might face when confronting fairness evaluation for the first time in a real recommendation setting.

To create these prototypes, we conducted a literature review to inform the creation of our decision tree prototype, particularly its high-level metric category leaf nodes. Our literature review formed the foundation for designing a process needed to scope a quantitative fairness context and identify a fairness metric category for said context. We created our original corpus by searching on the Google Scholar repository using the keywords "fairness," "ranking," "recommendation," "recsys," "fair," and "metric." We also added to our paper repository through snowball sampling; when we encountered a citation in one paper that referenced fairness metrics or definitions introduced in another paper, we added it to our repository. This process generated a corpus of 42 papers. We filtered this initial corpus to only include papers that introduced a new fairness metric. This removed papers that introduced re-ranking algorithms, or used previously defined fairness metrics. We also excluded papers that made assumptions that did not apply within practical large-scale recommendation contexts (e.g., if they assumed the practitioner had access to "ground truth" ranking labels) due to the motivation for this work to be useful in practical large-scale industry settings. This filtering process resulted in a total of 24 papers for the creation of our decision tree. Fourteen of these papers introduced provider fairness metrics, while nine of these papers introduced consumer fairness metrics. One paper introduced both provider and consumer metrics. It is important to note that this literature review should not serve as a comprehensive literature review since its goal was to scope metric *categories* for our prototype, not uncover every possible fairness metric. A list of these final papers we leveraged for the creation of our prototype can be found in the "Literature Review" section after the References section.

3.1.1 Scoping harms for Interviews. Before interviewing participants, we asked them to complete an online questionnaire to begin thinking about which algorithmic harms they might want to explore during the interview. The questionnaire first asked participants to select a familiar recommendation or ranking system to evaluate. Next, participants were prompted to scope potential harms. We used these as the basis for navigating through the decision tree during their interview to scope quantifiable fairness objectives and identify their related metrics. Participants were also asked to answer questions about their current role, their connection with the system they had chosen, and their comfort and knowledge of fairness metrics. Note that these interviews were purposely framed to focus on fairness-related algorithmic harms rather than positive impacts, to gain better insight into how to support practitioners that are asked to evaluate commercial systems to mitigate potential negative outcomes.

3.1.2 Prototyping the Decision Tree. Our original designs of the decision tree attempted to confront some of the challenges that previous research had uncovered regarding scoping quantitative fairness definitions. For the decision tree, we wanted to help practitioners scope their pre-defined qualitative harm formulated from the questionnaire into a quantitative harm with respect to common fairness terminology. To do this, we designed various "decision-making nodes" with corresponding branches (decisions) to demonstrate the differences between constraints such as provider versus consumer fairness, individual versus group fairness, and ranking versus distributional fairness. These branches were designed as a way to organize the content we found during our literature review, where each branch corresponds with a related fairness constraint that we encountered in the literature. This decision tree also accounts for the fairness constraints we described in Section 2.

The first decision-making node in the tree confronts the multi-stakeholder constraint, allowing practitioners to choose a branch between evaluating fairness for providers or consumers. The next decision node confronts the individual versus group constraint for both provider and consumer branches. For the providers branch, more decision-making nodes allow practitioners to select if they would like to measure fairness between multiple groups or for one group at a time. This design decision reflected current literature on provider group fairness, which introduces more nuanced measurements concerning groups.

The final decision-making nodes address system component (e.g., measuring fairness in item distributions versus in item rank positions) and fairness measurement property constraints found in our literature review. We do not designate a system component constraint branch for scoping consumer or individual provider fairness definitions due to the lack of literature for those specific paths. The leaf nodes represent specific fairness objectives addressing the various constraints. For example, when navigating through the decision tree for a multi-group provider ranking fairness context, the final node might ask the practitioner to choose between two

fairness objectives: (1) does the top-k ranking contain a specific proportion of items from these provider groups? or (2) are clicks, exposure, ranking, etc., proportional to item relevance for these provider groups? This choice of fairness objective leads the practitioner to a leaf node. Each leaf node has an associated "fairness category" that includes a list of metrics that fit within that category, both of which can be explored further in an associated spreadsheet. Table 2 provides an overview of these fairness categories and their associated constraints, objectives, and metrics. The final design of the decision tree included in this paper is the result of five iterations during our interviews, and iteration while drafting this final paper to facilitate follow up research into wider use beyond the study's specific organization setting. Note that this is not meant as a final usable 'product', but rather as a starting point for more research and design. This version of the decision tree can be found in the Appendix.

3.2 Interviews

We conducted a total of 15 semi-structured interviews with machine learning (ML) or ML-adjacent practitioners working in content recommendation in one organization. Each participant provided informed consent to participate in the study. All participants were working in a commercial consumer-provider recommender system setting with millions of users worldwide. The participants referenced in this paper can be found in Table 1.

All interviews were conducted online via Google Meet, and participants were not additionally compensated for their voluntary participation. Rutakumwa et al. [45] showcased that opting to not record interviews while maintaining robust written documentation can preserve participant privacy while ensure results do not change. As such, our interviews were not recorded to allow for participant privacy; all quotes from participants were gathered from notes taken during interviews. Participants were recruited via snowball sampling, where previous participants connected the research team with other relevant practitioners who they had worked with who might be a good fit for the study. The research team conducted inductive, thematic analysis [11] on interview notes, and categorized participant responses and observations into themes and sub-themes, which we detail in the following section.

4 RESULTS & DISCUSSION

In this section, we describe the challenges explicitly mentioned or implicitly observed while practitioners attempted to scope fairness-related conceptualizations for their selected recommender system. We compare these challenges to those identified in previous research and explore the implications of our results.

4.1 Scoping Qualitative Harms

Research literature popularly defines impact in terms of "fairness", but we found that this wording may not adequately reflect the various impacts practitioners need to evaluate in an industry setting. This finding was uncovered while discussing the results of the scoping worksheet with each participant to understand which potential impact they would like to use while navigating the decision tree. We found that some had encountered challenges when deciphering the ambiguity of the term "fairness" while scoping their fairness problem. Interview participants expressed having difficulty knowing how to define "fairly" in a quantifiable way. P14 described "fairly" as *"legalese,"* , noting that we would need a concrete definition of fairness, such as *"in proportion to the opportunity available to [providers],"* in order to capture this concept accurately. P13 similarly described that they would need a definition for "fair," ideally based on market research and in line with company values and business goals. This can include a conceptualization of, for example, concrete algorithmic impacts rather than 'algorithmic fairness'. This suggests that it could be helpful for practitioners to relate the choice of impact conceptualization, as well as metrics, to specific company and stakeholder goals or values, which has also been discussed in previous work [13, 33]. This also points to the need for dedicated business support, which is often more feasable in larger corporate settings, leaving smaller organizations at a disadvantage due to less access to resources [3].

4.2 Navigating the Decision Tree

Here we share the challenges participants faced when navigating between decision-making nodes and branches in the decision tree.

4.2.1 The Multistakeholder Constraint Branch. Beyond scoping qualitative harms and their associated fairness objectives, as noted above, participants struggled to understand common nuances used in fairness research literature to define various fairness constraints. When evaluating the multistakeholder constraint branch, some participants were unsure if they should explore fairness metrics for providers or consumers. Even though the decision tree offered examples of each stakeholder (e.g., providers could be content creators, consumers could be end-users), some of the practitioners remained unsure of which stakeholder they should prioritize for measuring fairness. Another example of this difficulty in choosing between stakeholders arose during P5's interview, who indicated a potential challenge determining if a system's recommendations reinforced country stereotypes. During this interview, P5 assumed that a consumer fairness metric would be most appropriate, since stereotypes might immediately impact those who consume the recommended content. However, they also noted that provider-fairness metrics might be equally appropriate for their context, especially if providers were being stereotyped due to their country of origin. P4 experienced the same challenge. They noted that fairness concerns could impact both providers and consumers simultaneously because they are *"two sides of the content ecosystem"* (P4). This challenge aligns with previous work [33] that highlighted difficulties practitioners experience when deciding which stakeholder population to prioritize for fairness interventions in a multi-stakeholder environment. This suggests that practitioners may need more guidance when selecting which stakeholders to prioritize when conducting fairness interventions, particularly in scenarios where design trade-offs may be necessary [9].

4.2.2 The Group vs. Individual Constraint Branch. Participants also had difficulty navigating the branch addressing the individual and group fairness constraint. Interestingly, a lack of knowledge about the difference between individual and group fairness was sometimes accompanied by a heightened level of confidence about this choice. For example, P2, P5, P8, and P14 all reported that one of the easiest

Table 1: Participants from the interview study and their roles. The "Participant ID" column corresponds to the alias that we use to identify each participant throughout the paper.

Work Area	Roles	Participant ID
Machine Learning	ML Engineer, ML Engineer Manager, Product Manager	P1, P2, P5, P6, P7, P8, P3, P13
Data Science	Data Scientist, Data Science Manager	P4, P9, P10, P11
Research	Research Scientist, UX Researcher	P12, P14

decisions for them to make during the interview was whether they were measuring group or individual fairness—even though *all* of these participants chose an option for their specific fairness context that appeared less aligned with literature definitions. We categorized these participants' "more aligned" options as metrics that matched the population of users they had chosen for their context. For example, P5 specified that they wanted to measure the impact of stereotypes on providers between different countries. In this case, it would be more "aligned" to define the provider country as a group type and measure group fairness. However, P5 chose to measure individual fairness instead.

For fairness problems that aligned with group provider fairness metrics, the decision tree prompted participants to choose whether they were interested in measuring fairness for one group at a time or multiple groups at a time. This decision also appeared challenging for some participants. For many fairness problems, both kinds of group fairness metrics could be applicable. For example, P8 wanted to measure whether certain long-tail creators were being unfairly represented in recommendation lists. This participant described that they could measure fairness by comparing their long-tail artist group's exposure against a target distribution (one group at a time) or by comparing their long-tail artist group's exposure against another long-tail artist group (multiple groups at a time). P8 said that both of these approaches might be equally appropriate, which made it difficult for them to know which branch to choose. This experience again reaffirmed our suspicion that there may be multiple, equally appropriate metric categories for a given fairness problem; which might add more difficulty for practitioners when deciding which of these many metric(s) to implement.

4.2.3 System Component Branch. The challenge of discerning nuances in fairness measurement terminology persisted when practitioners were asked to choose between system components for measurement. This branch splits based on the need for ranking or distribution metrics (e.g., measuring fairness of item exposure across all recommendation lists versus measuring fairness of item exposure in the top-k ranking of a specific recommendation list). At this point, the framework asks practitioners if they are more concerned with "poor representation in any group's item distribution or candidate generation list," (distribution metrics) or if they are more concerned with "any group being low in rankings or engagement for top-k or candidate generation lists," (ranking metrics).

P4, P10, P13, and P14 all stated that this choice between distribution and ranking fairness was the hardest decision they had to make in the decision tree. Both P4 and P13 were unsure which option to choose, since their fairness problem could apply to both categories. P4, P10, and P14 shared that this decision was difficult as they felt they did not have the specialized knowledge they needed concerning metrics that would fit their recommendation system. For example, these three participants disclosed that they did not understand some of the more technical vocabulary used to describe distribution versus ranking metrics. This suggests that education efforts might be best focused to target the gaps in practitioners knowledge: for some education about technical jargon and system components; for others education about fairness terminology.

4.2.4 The Fairness Objective Constraint. When participants did not have the necessary knowledge to make an informed decision on which fairness objective to finally choose, they sometimes chose the "easiest" path. For example, one participant navigated through the decision tree by selecting the fairness options and paths that were most familiar to them, even if they might have been missing out on metrics that were better aligned with their given scenario. This participant further described that the "easiest path" for them was one that tried to optimize "utility" in the system, because that path aligned with terminology also used for business goals, and it *"would be easier to get PM buy-in if needed."* This suggests that practitioners may choose metrics that are easy to understand, implement, or explain using existing business terms or perspectives, which may be the main priority to move the needle. Previous work has also described this phenomenon, where standard or well-known fairness definitions are adopted and assumed to be applicable across contexts [48]. Note that such considerations are valid—a very complex metric that is hard to understand but technically correct is unlikely to change decision-making, however, it is essential to clarify the consequences of such choices to assess whether a chosen metric is an appropriate proxy. This challenge could also be attributed to our choice of binary options at several of our decision-making nodes. Future tool iterations may benefit from allowing users greater flexibility in choosing from multiple possibilities.

4.3 Challenges Beyond the Framework

Throughout our interviews, we observed that there is not always a "correct" path or metric. Rather, there might be many equally-appropriate ways to measure fairness for a given scenario, all of which target evaluating different metric categories. When encountering these challenges, we asked participants to describe *who* they would ask for decision-making guidance. We found that participants were uncomfortable making fairness-related decisions on their own, especially when they faced challenges deciding on constraints.

Some of our participants expressed a desire to involve a diverse group of roles in these types of conversations, such as domain experts, model owners, product managers, engineers, user researchers, and business leaders. This was shown by P10 and P13, who both wanted to involve a *"fairness expert"* to help them feel confident in navigating this area. Beyond a fairness expert, P13 wanted to

Table 2: Fairness metrics with their associated categories and constraints as depicted in the decision tree. Some fairness metrics were assigned names if they were not explicitly named in the associated paper.

Stakeholder	Fairness Category	Fairness Constraints & Objectives	Metric(s) and References
Consumer	Individual Callibration	Recommendations should match consumers' previous interests.	Calibrated Fairness [18]
	Utility vs. Merit	System should distribute utility for consumers (individual or groups) based on their merit or need.	Generalized Cross Entropy [6]
	Group Utility	System should distribute utility equally between groups of consumers.	Non-Demographic Fairness [10], Average Accuracy [8], Differential Satisfaction [15], Pairwise Accuracy [3], epsilon Fairness [12]
	Group Error	System should have similar error rates between groups of consumers.	Value Unfairness [21], Absolute Unfairness [21], Underestimation Unfairness [21], Non-Demographic Fairness [10]
	Group Identity	Recommended items for consumers should not be related to group identity.	Non-Parity Unfairness [21], Discriminatory Skew [1], Pairwise Accuracy [3], Counterfactual Fairness [13]
Provider	One-Group Representation	Provider group representation in recommendations should match a pre-defined baseline distribution	Minimax KL Divergence [5], Average Provider Coverage Rate [14]
	One-Group Rank Utility	Groups of items should be ranked according to their utility for consumers.	Equal Expected Exposure [7], Inter-group Pairwise Accuracy [3]
	One-Group Rank Proportions	Top-k ranking should contain a specific proportion of items from a specific provider group.	Skew at K [9], Top-K Fairness [2], Normalized Discounted Difference [20], Viable-Lambda Test [16]
	Multi-group Item Exposure	Provider groups should have an equal item exposure distribution.	Minimax KL Divergence [5]
	Multi-group Item Relevance	Provider groups' probability distributions, preference ratings, exposure, etc. should be proportional to their items' relevance.	Kolmogorov-Smirnov Statistic [24], Absolute Difference in Mean Ratings [24]
	Multi-group Rank Proportion	Top-k ranking should contain a specific proportion of items from provider groups.	Normalized Discounted KL-Divergence [20], Normalized Discounted Ratio [20], Viable-Lambda Test [16], Ranked Group Fairness [22], Rank Parity [11]
	Multi-group Rank vs. Relevance	Provider groups item rankings should be equally proportional to their items' relevance.	Disparate Impact Ratio [17], Demographic Parity Constraint [17] , Disparate Treatment Ratio [17] , Listwise Fairness [22] , Inter-group Pairwise Fairness [3]
	Under-Over Exposure	Provider groups and/or items should not be systematically under-exposed or over-exposed in recommendations.	Gini Index [19]
	Individual Rank vs. Relevance	Item clicks, exposure, ranking, etc. should be proportional to item relevance for similar items.	Equal Expected Exposure [7], Disparate Treatment Ratio [17], Disparate Impact Ratio [17], Demographic Parity Constraint [17], Listwise Fairness [23], Equitable Individual Amortized Attention [4]

involve a user researcher to help them refine the population for fairness analysis as well as business leaders to help them align fairness tasks with business goals. P14, a user researcher, wanted guidance from more quantitative data-oriented counterparts such as data engineers, data scientists, and analytics engineers in order to understand *"the limitations of the model [because] they know what the data that we collect is actually capable of capturing"* (P14). P13 also further described that even though it would be nice to make these decisions with a group of fairness experts, they also did not want to exclude anyone who had little knowledge of fairness but was still interested in participating in the decision-making process.

4.3.1 Decision-Making and Implementation Gap. During our interviews, some practitioners who focused more on implementation indicated a need to involve other disciplines for fairness decisions in theory, but were less optimistic about the extent to which colleagues could be helpful in practice. For example, P9 described that these scoping decisions should ideally involve a combination of individual engineers or product developers working with a PM. However, in practice, P9 found that even though the PM should be involved, in a prior workplace they had experienced that fairness decisions ended up coming down to individual engineers implementing metrics. P5 had a similar experience to this. They described that from their perspective, ideally, fairness rationale and a more specific problem statement would come from a PM, but more detail would be necessary to implement than what would likely be provided. This points to the gap between conceptualization versus more detailed choices necessitated during implementation and needed clarity in roles and expectations in this process.

As seen with many other participants, P7 (who is an ML-focused engineer) was struggling to choose between the distributional and ranking metric constraints, and stated that part of their difficulty making this choice was because *"this is not just an ML problem, but also a product decision"* (P7). They said that in practice, they would probably go to their PM to help them refine their fairness goals, and then would bring the fairness scenario up with their entire engineering team. Then, they would probably try prototyping some of the metrics, and would check their results with a data scientist. They described the whole process as, *"a lot of different decisions made by different people, together and separately"* (P7). This description of fairness work as a collaborative, interdisciplinary process has also been explored in [39, 40], where the authors describe how decisions in practice are often made by a combination of different actors, such as data scientists, PMs, and user researchers, and are decided through collaborations and negotiations. Similar to that previous work, both P7 and P10 described their ideal collaborative process as one that highlights everyone's area of expertise.

These participants described that PMs would be the most helpful for high-level fairness decisions (e.g., the stakeholder constraint or the fairness objective constraint), while people like ML engineers could be responsible for the more low-level implementation decisions (e.g., the system component constraint). Previous work has also highlighted that practitioners are uncertain about *who* should be doing ethics work, and where the responsibility for doing ethics work should fall within an organization [35]. Organizations might benefit from providing central guidance, including clarity on roles and responsibilities, while future academic work or case studies could explore the consequences of organizational choices.

5 LIMITATIONS

Although the results of this work have broader implications for the WWW community, our research has limitations that could impact its generalizability. First, our interview study was conducted on 15 participants at one organization for development of domain-specific decision-making support. Though previous work has shown that small sample sizes can yield useful results for formative HCI research [10], our results still reflect the perspectives of those participants and might not generalize to other practitioners. Additionally, while we tried to recruit a diverse group of roles, the majority of our participants were machine learning engineers or data scientists. Future work could explore additional perspectives from product decision-makers or project and product managers. Finally, our decision-making framework was iteratively designed in a specific context, and might not take into account constraints in other domains.

Since we designed our framework to act as an exploratory case-study, no empirical evaluation has been conducted on the final iteration, and there could be metrics or constraints missing from the decision-tree. We purposely focused on metrics that authors conceptually positioned as 'fairness'-related. Expansions including other framings, such as 'disparate algorithmic impact,' would be helpful to be able to contrast additional available metrics. We recommend that future work refines these tools and framings, and investigates additional metrics or constraints in different domains.

6 CONCLUSION & FUTURE WORK

Scoping and identifying fairness metrics for recommender systems in practice is a complex task that requires guidance. Through an extensive literature review, framework design, and semi-structured interviews, we explored which challenges practitioners face when scoping a qualitative construct to a quantitative measurement for a specific context. We detail iterative requirements that may emerge when offering decision support that outlines concrete metric constraints and characteristics. We found that this type of guidance can also clarify which additional information practitioners need when formalizing their fairness definitions and goals in a way that aligns with their organization's values and in determining which metric constraints match those goals. For example, it is also important to accompany metrics guidance with guidance on *who* should be making decisions in practice, which support is available, and how education or tooling can cater to an audience with a variety of backgrounds—even within the same organization. Questions that occur while navigating different metric categories can help highlight what information is already known, or where process support may be necessary.

To alleviate some of these barriers, tooling and education efforts are crucial to support practitioners. Solutions for these challenges are not necessarily straightforward, as they require striking a balance between standardizing processes while also situating standards within the unique context of each system and its associated platform; this tension in standardizing measurement and processes in practice is a great candidate to explore in future work. We also recommend that future research focuses on working with industry practitioners and live recommendation systems to understand the real-world needs and obstacles practitioners face when incorporating algorithmic impact or fairness metrics into their workflows and systems. Ideally, such practical collaborations could help lower the barrier to implementing literature-proposed metrics so they are more available to create real-world impact. We hope this work can inspire future research directions to help practitioners evaluate approaches to algorithmic impact assessment to accurately capture the real experiences of people that encounter their systems and the practical constraints of online ranking and recommendation systems infrastructure.

REFERENCES

[1] Julia Angwin and Terry Jr. Parris. 2016. Facebook lets advertisers exclude users by race. https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race

[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

[3] Lex Beattie, Dan Taber, and Henriette Cramer. 2022. Challenges in Translating Research to Practice for Evaluating Fairness and Bias in Recommendation Systems. In *Proceedings of the 16th ACM Conference on Recommender Systems*. 528–530.

[4] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

[5] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in Recommendation Ranking through Pairwise Comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2212–2220. https://doi.org/10.1145/3292500.3330745

[6] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 514–524.

[7] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).

[8] Amanda Bower, Kristian Lum, Tomo Lazovich, Kyra Yee, and Luca Belli. 2022. Random Isn't Always Fair: Candidate Set Imbalance and Exposure Inequality in Recommender Systems. (2022). https://doi.org/10.48550/ARXIV.2209.05000

[9] Robin Burke. 2017. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093* (2017).

[10] Kelly Caine. 2016. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 981–992.

[11] Victoria Clarke, Virginia Braun, and Nikki Hayfield. 2015. Thematic analysis. *Qualitative psychology: A practical guide to research methods* 3 (2015), 222–248.

[12] Henriette Cramer, Jean Garcia-Gathright, Aaron Springer, and Sravana Reddy. 2018. Assessing and addressing algorithmic bias in practice. *Interactions* 25, 6 (2018), 58–63.

[13] Henriette Cramer, Jean Garcia-Gathright, Aaron Springer, and Sravana Reddy. 2018. Assessing and Addressing Algorithmic Bias in Practice. *Interactions* 25, 6 (oct 2018), 58–63. https://doi.org/10.1145/3278156

[14] Kate Crawford. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*.

[15] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. 2022. A Survey of Research on Fair Recommender Systems. *arXiv preprint arXiv:2205.11127* (2022).

[16] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. *arXiv preprint arXiv:2205.06922* (2022).

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

[18] Michael D Ekstrand, Anubrata Das, Robin Burke, Fernando Diaz, et al. 2022. Fairness in Information Access Systems. *Foundations and Trends® in Information Retrieval* 16, 1-2 (2022), 1–177.

[19] Ben Green and Lily Hu. 2018. The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Proceedings of the machine learning: the debates workshop*.

[20] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.

[21] HRW. 2018. *" Only Men Need Apply": Gender Discrimination in Job Advertisements in China*. Human Rights Watch.

[22] Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 375–385.

[23] Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. 2019. Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 383–390.

[24] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[25] Caitlin Kuhlman, Walter Gerych, and Elke Rundensteiner. 2021. Measuring group advantage: A comparative study of fair ranking metrics. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 674–682.

[26] Caitlin Kuhlman, MaryAnn VanValkenburg, and Elke Rundensteiner. 2019. FARE: Diagnostics for Fair Ranking Using Pairwise Error Metrics. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 2936–2942. https://doi.org/10.1145/3308558.3313443

[27] Po-Ming Law, Sana Malik, Fan Du, and Moumita Sinha. 2020. Designing Tools for Semi-Automated Detection of Machine Learning Biases: An Interview Study. *arXiv preprint arXiv:2003.07680* (2020).

[28] Tomo Lazovich, Luca Belli, Aaron Gonzales, Amanda Bower, Uthaipon Tantipongpipat, Kristian Lum, Ferenc Huszar, and Rumman Chowdhury. 2022. Measuring disparate outcomes of content recommendation algorithms with distributional inequality metrics. *arXiv preprint arXiv:2202.01615* (2022).

[29] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 205395171875668 4.

[30] Michelle Seng Ah Lee, Luciano Floridi, and Jatinder Singh. 2020. From fairness metrics to key ethics indicators (keis): a context-aware approach to algorithmic ethics in an unequal society. *Available at SSRN* (2020).

[31] Michelle Seng Ah Lee and Jat Singh. 2021. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.

[32] Weiwen Liu, Jun Guo, Nasim Sonboli, Robin Burke, and Shengyu Zhang. 2019. Personalized fairness-aware re-ranking for microlending. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 467–471.

[33] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–26.

[34] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[35] Jacob Metcalf, Emanuel Moss, et al. 2019. Owning ethics: Corporate logics, silicon valley, and the institutionalization of ethics. *Social Research: An International Quarterly* 86, 2 (2019), 449–476.

[36] Smitha Milli, Luca Belli, and Moritz Hardt. 2021. From optimizing engagement to measuring value. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 714–722.

[37] Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. 2021. Assembling accountability: algorithmic impact assessment for the public interest. *Available at SSRN 3877437* (2021).

[38] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, Vol. 1170. 3.

[39] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. In *Proceedings of the conference on fairness, accountability, and transparency*. 39–48.

[40] Samir Passi and Steven J Jackson. 2018. Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–28.

[41] Gourab K Patro, Lorenzo Porcaro, Laura Mitchell, Qiuyue Zhang, Meike Zehlike, and Nikhil Garg. 2022. Fair ranking: a critical review, challenges, and future directions. *arXiv preprint arXiv:2201.12662* (2022).

[42] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.

[43] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. 2020. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 131–141.

[44] Brianna Richardson, Jean Garcia-Gathright, Samuel F Way, Jennifer Thom, and Henriette Cramer. 2021. Towards Fairness in Practice: A Practitioner-Oriented Rubric for Evaluating Fair ML Toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[45] Rwamahe Rutakumwa, Joseph Okello Mugisha, Sarah Bernays, Elizabeth Kabunga, Grace Tumwekwase, Martin Mbonye, and Janet Seeley. 2020. Conducting in-depth interviews with and without voice recorders: a comparative analysis. *Qualitative Research* 20, 5 (2020), 565–581.

[46] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577* (2018).

[47] Jeffrey Saltz, Michael Skirpan, Casey Fiesler, Micha Gorelick, Tom Yeh, Robert Heckman, Neil Dewar, and Nathan Beard. 2019. Integrating ethics within machine learning courses. *ACM Transactions on Computing Education (TOCE)* 19, 4 (2019), 1–26.

[48] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.

[49] Jessie Smith, Nasim Sonboli, Casey Fiesler, and Robin Burke. 2020. Exploring user opinions of fairness in recommender systems. *arXiv preprint arXiv:2003.06461* (2020).

[50] Nasim Sonboli, Robin Burke, Zijun Liu, and Masoud Mansoury. 2020. Fairness-aware Recommendation with librec-auto. In *Fourteenth ACM Conference on Recommender Systems*. 594–596.
[51] Nasim Sonboli, Farzad Eskandanian, Robin Burke, Weiwen Liu, and Bamshad Mobasher. 2020. Opportunistic multi-aspect fairness through personalized re-ranking. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 239–247.
[52] Michael Veale and Reuben Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society* 4, 2 (2017), 2053951717743530.
[53] Sahil Verma, Ruoyuan Gao, and Chirag Shah. 2020. Facets of fairness in search and recommendation. In *International Workshop on Algorithmic Bias in Search and Recommendation*. Springer, 1–11.
[54] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review* 41 (2021), 105567.
[55] C Xu and T Doshi. 2019. Fairness indicators: scalable infrastructure for fair ML system. *Mountain View (CA): Google (accessed 2020-01-27). https://ai. googleblog. com/2019/12/fairness-indicators-scalable. html* (2019).
[56] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1569–1578.

LITERATURE REVIEW

[Lit1] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 199 (nov 2019), 30 pages. https://doi.org/10.1145/3359301
[Lit2] Abolfazl Asudeh, HV Jagadish, Julia Stoyanovich, and Gautam Das. 2019. Designing fair ranking schemes. In *Proceedings of the 2019 international conference on management of data*. 1259–1276.
[Lit3] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in Recommendation Ranking through Pairwise Comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2212–2220. https://doi.org/10.1145/3292500.3330745
[Lit4] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 405–414. https://doi.org/10.1145/3209978.3210063
[Lit5] Anubrata Das and Matthew Lease. 2019. A conceptual framework for evaluating fairness in search. *arXiv preprint arXiv:1907.09328* (2019).
[Lit6] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. 2022. A Survey of Research on Fair Recommender Systems. *arXiv preprint arXiv:2205.11127* (2022).
[Lit7] Fernando Diaz, Bhaskar Mitra, Michael D. Ekstrand, Asia J. Biega, and Ben Carterette. 2020. Evaluating Stochastic Rankings with Expected Exposure. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 275–284. https://doi.org/10.1145/3340531.3411962
[Lit8] Michael D Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Conference on fairness, accountability and transparency*. PMLR, 172–186.
[Lit9] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*. 2221–2231.
[Lit10] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*. PMLR, 1929–1938.
[Lit11] Caitlin Kuhlman, MaryAnn VanValkenburg, and Elke Rundensteiner. 2019. FARE: Diagnostics for Fair Ranking Using Pairwise Error Metrics. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 2936–2942. https://doi.org/10.1145/3308558.3313443
[Lit12] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-Oriented Fairness in Recommendation. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 624–632. https://doi.org/10.1145/3442381.3449866
[Lit13] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. 2021. Towards personalized fairness based on causal notion. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1054–1063.
[Lit14] Weiwen Liu, Jun Guo, Nasim Sonboli, Robin Burke, and Shengyu Zhang. 2019. Personalized fairness-aware re-ranking for microlending. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 467–471.
[Lit15] Rishabh Mehrotra, Ashton Anderson, Fernando Diaz, Amit Sharma, Hanna Wallach, and Emine Yilmaz. 2017. Auditing Search Engines for Differential Satisfaction Across Demographics. In *Proceedings of the 26th International Conference on World Wide Web Companion* (Perth, Australia) *(WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 626–633. https://doi.org/10.1145/3041021.3054197
[Lit16] Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the Impact of User Attentionon Fair Group Representation in Ranked Lists. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 553–562. https://doi.org/10.1145/3308560.3317595
[Lit17] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 2219–2228. https://doi.org/10.1145/3219819.3220088
[Lit18] Harald Steck. 2018. Calibrated Recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 154–162. https://doi.org/10.1145/3240323.3240372
[Lit19] Saúl Vargas and Pablo Castells. 2014. Improving sales diversity by recommending users to items. In *Proceedings of the 8th ACM Conference on Recommender systems*. 145–152.
[Lit20] Ke Yang and Julia Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management* (Chicago, IL, USA) *(SSDBM '17)*. Association for Computing Machinery, New York, NY, USA, Article 22, 6 pages. https://doi.org/10.1145/3085504.3085526
[Lit21] Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. *Advances in neural information processing systems* 30 (2017).
[Lit22] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1569–1578.
[Lit23] Meike Zehlike and Carlos Castillo. 2020. Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of The Web Conference 2020*. 2849–2855.
[Lit24] Ziwei Zhu, Xia Hu, and James Caverlee. 2018. Fairness-aware tensor-based recommendation. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 1153–1162.

Decision Tree Framework (Consumers)

***Publication Note*: This version of the Decision Tree is not meant as a final product. It is a draft tool for the research community to further explore metric options available and their constraints.**

Instructions

Begin with a specific harm or unfair impact use-case for your recommender, ranking, or search system, then follow the decision-tree to find your options for appropriate fairness metrics to measure that impact. Once you find your way to a "leaf" node labeled "METRIC CATEGORY," you can find more information about those metric(s) and their associated academic papers by referencing their Fairness Category in Table 2. You can zoom in/out and scroll on this page to see text more easily.

Please note that the metrics linked from this decision tree are presented as options for your fairness measurement based on what is commonly used in literature. They are not meant to be prescriptive nor all-encompassing and we recommend you discuss these choices with relevant stakeholders before deciding on your metric(s). There may be more than one appropriate metric category for your use-case!

Key

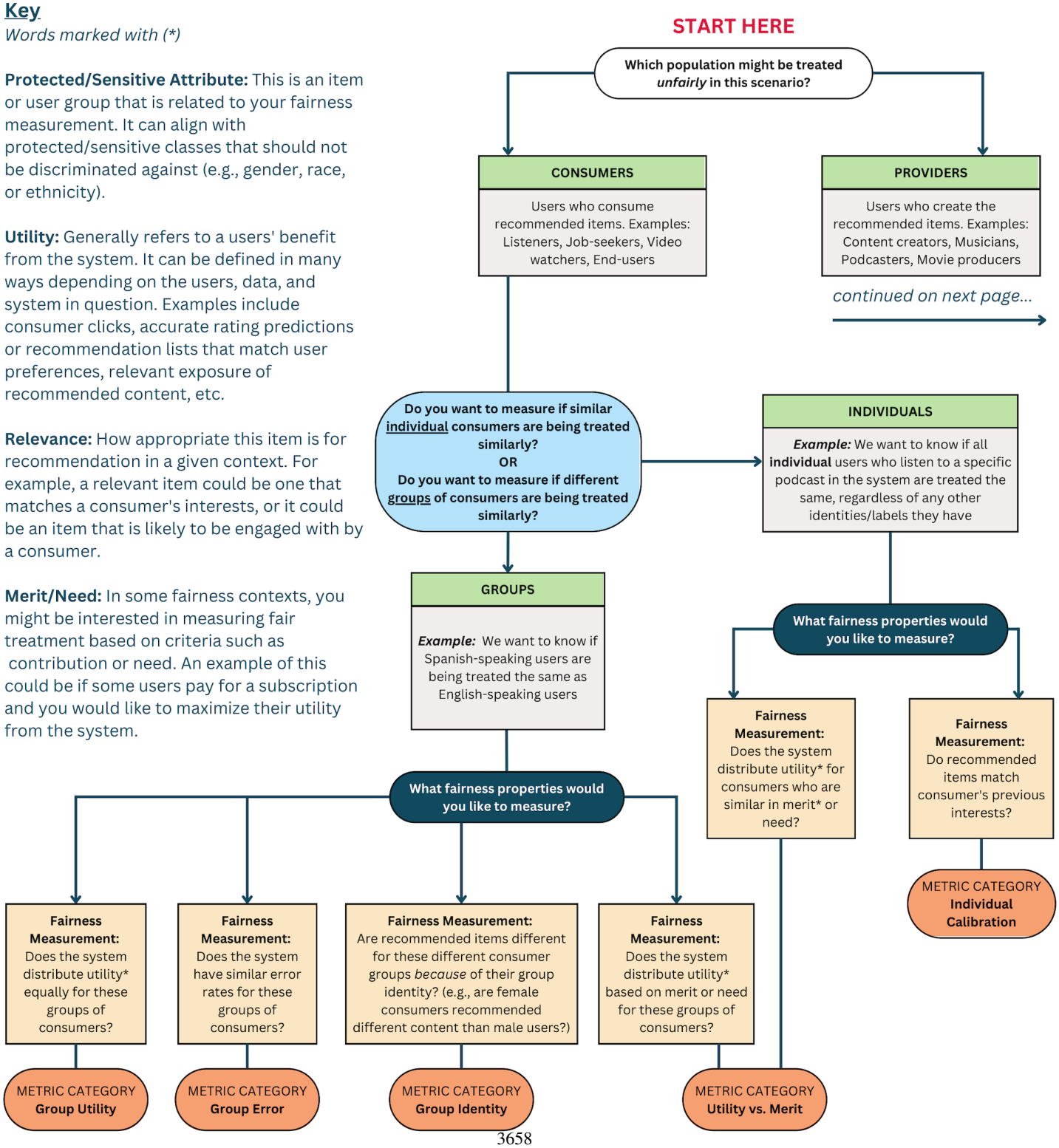
Words marked with (*)

Protected/Sensitive Attribute: This is an item or user group that is related to your fairness measurement. It can align with protected/sensitive classes that should not be discriminated against (e.g., gender, race, or ethnicity).

Utility: Generally refers to a users' benefit from the system. It can be defined in many ways depending on the users, data, and system in question. Examples include consumer clicks, accurate rating predictions or recommendation lists that match user preferences, relevant exposure of recommended content, etc.

Relevance: How appropriate this item is for recommendation in a given context. For example, a relevant item could be one that matches a consumer's interests, or it could be an item that is likely to be engaged with by a consumer.

Merit/Need: In some fairness contexts, you might be interested in measuring fair treatment based on criteria such as contribution or need. An example of this could be if some users pay for a subscription and you would like to maximize their utility from the system.



Decision Tree Framework (Providers)

Publication Note: This version of the Decision Tree is not meant as a final product. It is a draft tool for the research community to further explore metric options available and their constraints.