4.2 Fairness Evaluation

Christine Bauer (Paris Lodron University Salzburg – Salzburg, Austria, christine.bauer@plus.ac.at),
Michael Ekstrand (Drexel University – Philadelphia, USA, mdekstrand@drexel.edu),
Andrés Ferraro (SiriusXM, Spain, andresferraro@acm.org),
Maria Maistro (Copenhagen University – Denmark, mm@di.ku.dk)
Manel Slokom (CWI – The Netherlands, manel.slokom@cwi.nl),
Robin Verachtert (DPG Media, Belgium, robin.verachtert@dpgmedia.be)

This group focused on paradigms and practices for evaluating the fairness of a recommender system. As noted in Ekstrand's talk abstract (Section 3.5), fairness is a complex, nuanced, and context-dependent family of problems that defies simple definitions or overly-standardized evaluation approaches [20, 42]. It is, however, a vital problem: recommendation brings significant benefits to users, creators, and society by catalyzing economic opportunity and enabling effective access to a wider range of art, news, information, and products. Ensuring that these benefits accrue broadly across society, instead of being concentrated on the few or distributed in ways that replicate historical and ongoing discrimination, is essential if recommendation is to truly serve the public good.

Because fairness metrics and evaluation requirements are specific to particular applications, fairness problems, and goals [44, 21], it is difficult to present technical best practices such as particular metrics, data processing strategies, etc. Instead, we seek to describe "best meta-practices": ways of approaching the planning, execution, and reporting of fairness evaluations that will enable work to be rigorous – both socially and technically – and clearly communicated. In this section, we synthesize ideas from prior work on problems and approaches to fairness research [17, 18, 21, 44, 49] to which we refer the reader for further study, along with some fresh observations of our own.

Many of the ideas in this section are not specific to fairness [18]; all aspects of recommender system evaluation benefit from careful attention to the problem, justification of metrics and methods, and clear communication.

4.2.1 Landscape

Understanding fairness in recommender systems requires considering a complex ecosystem of various entities and interconnected concepts. In Fig. 1, we briefly overview the main concepts behind fairness. The entities involved include consumers, item providers, and subjects; multiple actors can be considered together under multisided fairness. Fairness problems also often divide into individual and group problems, regardless of the entities involved. Additionally, we describe the potential harm caused by unfairness and the temporal dimension of fairness.

For "Who"?

Fairness becomes a critical factor when recommender systems are deployed in settings where harmful discrimination may occur. We distinguish between different classes regarding "who" fairness might concern [1, 18]. *Consumer side fairness* or user side fairness ensures

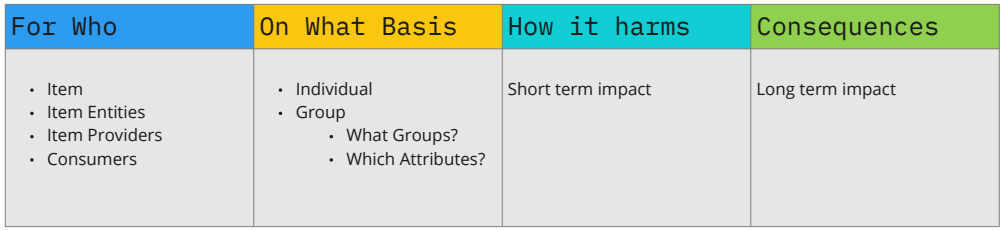

Figure 1 Categorization of fairness factors.

that consumers[7] of the recommender system are treated fairly in the quantitative and qualitative aspects of their experience. This involves ensuring equity of utility or usability, fair representation, avoiding stereotypes, etc. *Fairness towards item side entities* ensures a fair treatment of items; it can include provider and subject side fairness but can also be considered without knowledge of providers or subjects. A system can be unfair by treating similar items differently, e.g., when two news articles on the same topic and with comparable quality are not exposed equally. *Provider-side fairness* is an item-side entity concern which ensures fair treatment of item providers. *Subject-side fairness* is an item-side entity concern which ensures fair treatment of the subjects (people or entities) mentioned in, or related to the items. For example, in news recommendation, a system can be unfair to the gender of people described in news articles or to specific topics discussed in the articles. *Multisided fairness* [11] considers consumers and providers, demanding fairness on both sides.

On "What" basis?

Fairness is often characterized as individual vs. group fairness [17]. The goal of *individual fairness* is to treat similar individuals similarly, so that each individual receives an appropriate treatment in accordance with some task-specific notion of "merit". The goal of *group fairness* is to treat different groups similarly, so that there are no systematic disparities across groups. Usually, a protected group is contrasted against an unprotected group (also called dominant or majority group) to guarantee that protected individuals are treated comparably to unprotected ones. Groups are often defined upon attributes from anti-discrimination law, e.g., gender, ethnicity, religion, and age.

Individual fairness assumes a function to measure the similarity among individuals. Defining such similarity function is challenging due to the lack of ground truth, data biases, task dependency [25] and very often results in solving the task itself [12]. For example, a "perfect" similarity function based on user preferences and past interactions could be used to generate "perfect" recommendations. While group fairness might seem easier to accomplish, it requires access to protected attributes to define groups. These attributes are often unavailable or difficult to collect because they represent sensitive data, e.g., gender. Moreover, group fairness does not guarantee fair treatment among individuals within a group due to aggregation and comparison among groups (fairness gerrymandering [32]). For

[7] "Consumer" is commonly used to indicate the people using a recommender system. The term should not be used when the recommender system recommends people, such as in dating applications or job recommendations. For brevity and clarity, we will use consumer in this piece as we do not explicitly talk about these topics.

example, a music recommender system might achieve group fairness with respect to gender by increasing exposure for a single artist, but this does not ensure fairness for other artists of the same gender.

"How"?

Exploring the "How?" of fairness involves examining various dimensions through which fairness can be achieved or compromised. Here, we refer to some examples of how unfairness can lead to unfair distribution of utility, severe consequences, exposure, discrimination, misrepresentation, and reinforces stereotyping.

Unfair distribution of utility Unfairness in recommender systems can lead to unequal distribution of utility, where benefits such as opportunities are disproportionately allocated. When recommendations favor certain consumers/users or item providers over others due to biases in data or algorithms, some groups receive more exposure and advantages, while others are marginalized [22, 19, 24]. This inequitable distribution not only reduces the overall satisfaction and utility for disadvantaged users but also perpetuates existing inequalities and limits diversity.

- How can recommender systems be designed to ensure an equitable distribution of utility among all users/items/subjects?
- What factors contribute to the unfair distribution of utility in recommender systems?
- How do biases in the data and algorithms affect the distribution of utility among different user/item groups?
- What metrics can be used to measure the fairness of utility distribution in recommender systems?
- How can interventions be implemented to correct the unfair distribution of utility in existing recommender system algorithms?

Disparity of Exposure Depending on the user attention model that is considered, an item's position in the recommendation list determines the exposure of individuals or groups of items [7, 43]. Therefore, exposure has effects and implications on how much users will consume those individual or groups of items. Disparity of exposure is typically based on the principles of equality of opportunity. This can be further operationalized in different ways [15, 31].

For example, disparity of opportunity can be based on the idea that all item groups/similar items should get exposure proportional to the collective merit of the items in the group or the merit of individual items [30]. Fairness for individuals can be defined following the idea that exposure should be proportional to relevance for each subject in a system. In contrast, fairness for groups means that exposure should be equally distributed among members of groups defined by sensitive attributes such as gender and lyric language [43].

- How can exposure be measured and balanced to ensure fairness for all users and item providers?
- What algorithms or techniques can be used to ensure equitable exposure?
- How does unequal exposure affect user satisfaction and engagement with recommender systems?
- What are the challenges in achieving group-level exposure fairness, and how can they be addressed?
- How can exposure fairness be maintained over time as user preferences and content availability change?

Discrimination occurs when the algorithmic decisions tend to disadvantage certain groups based on characteristics such as demographic information, e.g., ethnicity, gender, age, or socioeconomic status [2].

- How does discrimination affect user trust and platform credibility?
- What are the legal and ethical implications of discrimination in recommender systems?
- How can inclusive data collection practices reduce the risk of discrimination in recommendations?

Misrepresentation refers to an inaccurate representation of users or item providers' characteristics [21, 17]. Misrepresentation can be in the form of inaccurately representing users' interests and information needs internally, preventing certain user groups from systematically having less accurate representations (e.g., user embeddings or other user models that may lead to stereotyped recommendations [21]. Providers can be harmed by not having their products consumed.

- How can misrepresentation in user profiles and item descriptions be identified and corrected in recommender systems?
- What impact does misrepresentation have on user satisfaction and item provider success?
- How do inaccurate user models contribute to the spread of stereotypes in recommendations?
- What techniques can improve the accuracy of user and item representations to prevent misrepresentation?
- How can transparency in recommender systems help users understand and correct potential misrepresentations?

Reinforcing stereotype refers to the potential of recommender system algorithms to perpetuate harmful or unnecessary stereotypes by consistently promoting content that aligns with narrow, stereotypical views [38].

- How do recommender systems contribute to the reinforcement of societal stereotypes?
- What are the long-term impacts of stereotype reinforcement on users and society?
- How can algorithms be designed to avoid reinforcing stereotypes?
- What role does diverse and inclusive data play in preventing stereotype reinforcement? How can user feedback be used to identify and mitigate the reinforcement of stereotypes in recommendations?

On "What" Scale?

Machine learning models often optimize some static objectives, causing fairness to be regarded as a static function. Most definitions consider fairness as a one-shot process, i.e., with respect to a single point in time. The underlying assumption is that fairness will be beneficial for the protected individuals or groups, as well as the whole society, in the long term. However, decisions based on ML models can be iterated over time, and one-step fairness can even cause harm [28, 34, 35, 13, 33, 6, 24].

Recommender systems are dynamic and interactive by nature, i.e., the nature of entities may change over time. For example, groups based on attributes such as popularity can quickly change over time, and fairness interventions can potentially drive items into or out of the top popular group. This distinction of fairness as a long-term or short-term process results in static vs. dynamic fairness. *Static fairness* disregards changes in the underlying environment, e.g., utility, attributes, etc., while *dynamic fairness* adapts to the environment [26].

The *severity of consequences* refers to the negative impact of unfair recommendations on all entities involved, e.g., consumers, item providers, etc. For instance, severe consequences for consumers can be in the form of missed opportunities, financial losses, or psychological harm. Item providers such as content creators or sellers can face severe consequences that manifest as reduced visibility and revenue (see Section 4.2.2 for concrete examples).

The extent to which unfair recommender systems can cause harm depends also on the temporal dimension. For example, disparity of exposure might not cause immediate harm but, if reiterated in the long-term, can potentially lead to severe discrimination, job and profit loss, and reinforcement of stereotypes. In the long term, unfairness can also have significant *societal consequences*. With news and social media sites, unfair recommender systems might promote content emphasizing only one political side or misinformation discriminating against certain groups [50, 5].

4.2.2 Examples / Use cases

Fairness concerns may be encountered in any recommender systems use case. Here, we present a few examples to give an intuition for what fairness concerns we might consider in research and practical applications. We chose two use cases to explore a subset of potential fairness concerns. By no means is this list exhaustive. More examples can be found in the literature available on this topic [17, 18, 49].

Research paper recommender system/search engine

Academic search and recommendation aim to help researchers find relevant papers for their interests. The widespread use of these systems calls for ways to ensure fair information access to avoid harmful consequences to authors, institutions, and journals. In Fig. 2, we briefly overview the main concepts behind fairness for the use case "research paper recommender systems".

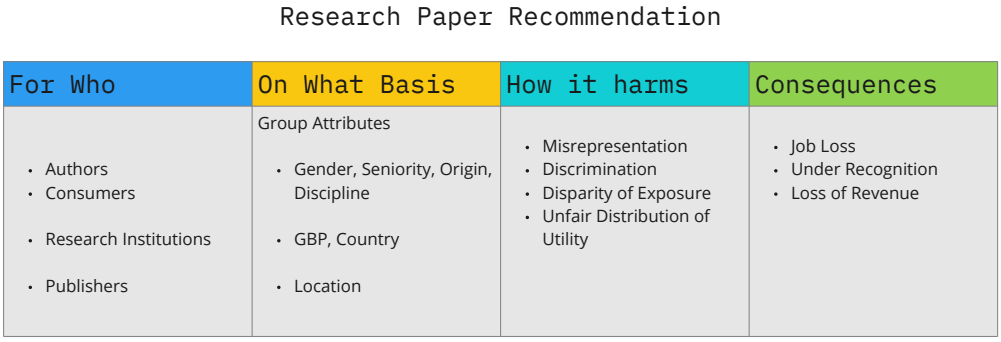

Figure 2 Identifying the key points of fairness in research paper recommender systems.

Possible actors involved are paper authors, users of the search or recommender system, research institutions, and publishing venues, e.g., conferences and journals. Author group fairness can be defined by attributes such as gender, seniority, geographical origin, or discipline. The Gross Domestic Product (GDP) and the country can apply to research institutions and country for publishers.

Examples of fairness concerns for this domain include:

- If the system provides an unfair disadvantage to a group of authors, this may lead to lower recognition in the field for this group of authors (discrimination, disparity of exposure, misrepresentation). Consequently, this can lead to challenges for them in finding a job posting in academia and a loss of revenue in the long term.
- If a discipline is under-represented, this can lead to a knowledge gap for the user (reader) of the system (disparity of exposure, misrepresentation). This knowledge gap can lead to less-informed papers and potential rejection of the work.
- If there is a systemic bias on the location or renown of an institution, this can lead to under-recognition for these institutions (discrimination, disparity of exposure, misrepresentation), thus stumping their growth, and harming their search for funding and students.
- If articles from a publisher or group of publishers are under-recommended (discrimination, disparity of exposure, misrepresentation), this can lead to a lower value for publications by this publisher and consequently to fewer submissions to the journal, leading to diminishing value for the publisher.

E-commerce

Online retailers provide users with easy access to products from all over the world. Online marketplaces such as Amazon, Zalando, and Ali-Express serve many users with products from various vendors. Thus, their recommender systems have an impact on the fairness towards many stakeholders. In Fig. 3, we briefly overview the main concepts behind fairness for the use case "e-commerce recommender systems".

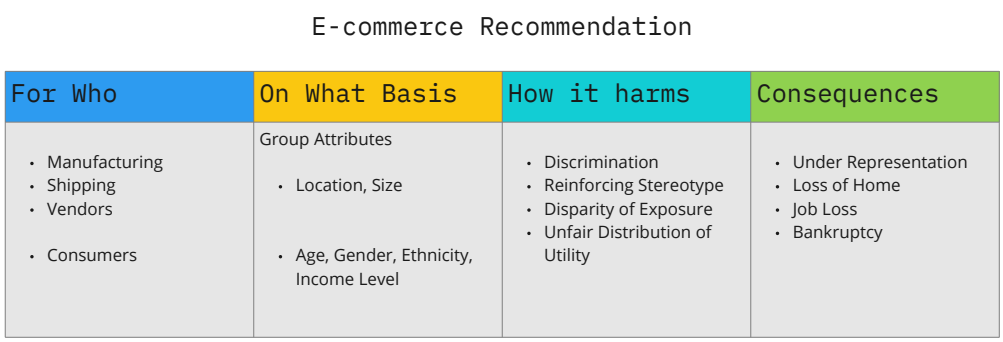

Figure 3 Identifying the key points of fairness in e-commerce recommender systems.

We identify two main classes of actors from the selling and buying side: companies involved in the production chain (manufacturer, vendor, shipping companies) and consumers. Meaningful attributes for companies are size and country. For consumers, we can consider gender, ethnicity, age group, and income level as relevant attributes.

Some specific concerns we would like to highlight are the following:

- If the system is under-recommending items from a group of vendors (discrimination, disparity of exposure, misrepresentation), this could lead to lower sales for these vendors. This, in turn, is likely to lead to a loss in revenue for them.
- If there is an unfair distribution of the manufacturing plants of recommended items, then underrepresented manufacturing plants might lose revenue as the items they make are not being sold as easily (discrimination, disparity of exposure, misrepresentation). This could lead to job loss for the employees and even bankruptcy.

- If one user group is consequently recommended more expensive items (discrimination, misrepresentation), this may lead to higher strains on their income; thus, introducing or reinforcing a monetary gap with the other groups.
- If recommendation quality is systemically lower for a group of users (unfair distribution of utility, misrepresentation), this leads to lower utility for them.
- If the recommender system consistently recommends stereotypical items to groups of users, this can lead to *reinforcing stereotypes*. For example, girls might get recommended books about princesses, while boys get books about knights.

4.2.3 Problem definition

As with any evaluation, for fairness, the problem to be evaluated has to be clearly defined [48]. In this regard, there are some specifics for fairness evaluation that we need to emphasize. First and foremost, a state of "full" fairness does not exist. Many dimensions come into play that might be considered unfair, but we can only know about it if we evaluate an RS on those dimensions. Thus, fairness evaluation needs to target a specific fairness problem and can only draw conclusions on this specific problem.

Depending on how we define the problem, a solution may be (un)fair with respect to that specific definition but not to another. Before describing the different aspects involved in defining the problem, it is important to highlight the connections and differences between fairness and bias. In general, the term "bias" may be used to refer to multiple concepts. [36] categorize biases as *statistical* or *societal*: 1) Statistical bias refers to the systematic differences between data or outputs and the underlying observable world; and 2) societal biases to the systematic differences between the observable world and the arguable ideal world without any form of discrimination. We use bias to describe the objective deviation or imbalance in a model, measure or data compared to an intended target, including both sampling biases and measurement error. Therefore, we use the term "bias" to refer to a **specific property or characteristic of the system without making any inherently normative judgment**. On the other hand, we use "fairness" to discuss the **normative aspects of the system and its effects.** Here, it is important to distinguish between the technical fact and the moral, ethical, or legal concern in the interests of societies as well as individuals.

Bias vs. fairness: Research on fairness in RSs can be of descriptive or normative nature, which will particularly shape the interpretation phase in the evaluation process. In its descriptive nature, the purpose of the evaluation of fairness aspects is to describe the current state (is situation) of one or several recommendation approaches in its given context (e.g., domain, dataset, constraints, assumptions). In a normative take on fairness, there is a target that should ideally be reached or approached (should-be situation). This may also include that different intervention strategies are evaluated for their effectiveness and compared accordingly (as, for instance, done in [24]). Note that there is not necessarily a specific target distribution or target figure on a particular metric to be targeted; instead, the goal is often a direction of how an intervention should compare to the is situation – thus "improvement" over the situation before (e.g., smaller gender gap than before, higher exposure of the minority group than before).

Context/Motivation: In the context of RSs, fairness-related harms arise when there is, for instance, an unequal distribution of utility (e.g., harming a fraction of users with specific probabilities). Accordingly, a fairness problem needs to be specified based on the specific harms that arise. As with any research problem, the fairness problem needs to be motivated based on prior research or real-life observations, underpinning the relevance of the harm. For

instance, [19] motivated the relevance of the investigated harm through previous research and practices on author gender aspects in the book domain. [24] conducted interviews with artists in the music domain to find out that this stakeholder group experiences particular harm due to gender imbalance, which was then the basis for motivating their RS fairness research on gender aspects (specifically, exposure of women) in this domain. When motivating and defining a fairness problem, it is crucial to care about an appropriate problem; specifically, *not* trivializing the problem into disrespect. Similarly, we need to be careful with "toy" problems: Is the problem causing harm? Should we give priority to researching this specific problem? Is it relevant in practice? Does it matter? In this regard, we need to contextualize the fairness problem: On the one hand, context is needed to motivate the relevance of the problem in its domain or more specific context (e.g., women and gender minorities are generally strongly underrepresented in the music domain [29], which contextualizes why and how artist gender fairness is addressed in this domain [24]). On the other hand, contextualization is needed for results interpretation (see Section 4.2.5).

Multiple definitions: The fairness problem we are working on can be defined in multiple ways. In the case of gender imbalance in music recommendation, female artists have less exposure than male artists since they are shown lower in the ranking; but also, there are fewer female artists recommended overall. Therefore, it is important to clearly define which aspect(s) the work is addressing. In order to do this, it is essential to take into account the context and motivation of the work: if the goal is to increase the consumption of female artists in the long term, increasing the number of female artists recommended could not be enough if they are consistently ranked lower than male artists [24]. Therefore, we need to ensure that the metric we use to measure and optimize our algorithm aligns with the specific dimension of fairness that we defined. For this, it is crucial to clearly define and document the research question that we are trying to address.

The multiple definitions are related to the high complexity of the problem we are working on. When defining the problem we want to address, we always need to make certain assumptions. For example, in the case of gender fairness, an assumption that authors make is that all artists in the dataset are annotated with a gender label [24]. This is an assumption that, in the real world, will either bring some limitations or require practitioners to find a way to operationalize that is out of scope in the proposed solution.

Multiple dimensions: The concept of multiple fairness dimensions means that there are multiple active concerns in a given system: gender, religion, sexual orientation, etc. When we define different groups of individuals that belong to more than one group, we need to consider a combination of the groups. Addressing multiple dimensions of fairness makes the problem more complex but also allows us to find issues that otherwise go unnoticed. For example, in the case of music recommendation, when promoting female artists to reach a more balanced consumption, it may happen that only female artists from Western countries are exposed but not from the Global South. Therefore, in this case, considering the multiple dimensions of fairness implies exposing, to some degree, female artists from both the Global North and the Global South.

To summarize, the fairness problem definition needs specificity in many regards:

- Specification of the harms/inequities that are being addressed; relevance and appropriateness need to be motivated
- Clear specifications of the fairness dimensions that are supposed to be addressed and evaluated
- Scoping and contextualization:
  - Clearly state the scope of the evaluation
  - Put the scope into context (different contextualization)

- Clearly explicate the assumptions
- Define scope, i.e., showing the existence or magnitude of a fairness issue (descriptive), investigating and evaluating fairness interventions
- Is the point of interest causality or correlation?

When defining the problem, it is helpful to keep the main concepts behind fairness in mind, as described in Section 4.2.1 (Fig. 1): Fairness "for who", "on what basis", "how it harms", and "consequences".

4.2.4 Operationalization & Planning

Defining the problem is only the beginning: once the problem is defined, it needs to be *operationalized* – i.e., translated into a specific evaluation design, including data set(s), method of running the experiment(s), and evaluation metric(s) [44, 21]. This operationalization process can result in qualitative, quantitative, or mixed-methods research designs.

This section briefly summarizes considerations for effectively operationalizing quantitative evaluations of recommender system fairness. We separate operationalization from the definition process to facilitate clearer thinking about the relationship between the specific measurements and the original social, ethical, policy, and technical goal(s). No one measurement can fully capture everything of interest, particularly for a concept as complex and multifaceted as fairness (even after defining a specific fairness problem), and it is vital to recognize and document what is missing in the specific evaluation design and avoid the trap of conflating the measurement with the original goal. [44], [21], and others provide further reading on scoping.

An effective evaluation design for fairness will have at least the following properties:

- It is **well-matched** to the particularities of the application and problem [21].
- It can be **effectively computed** with data that is available (or obtainable) and of high fidelity. In this regard, we emphasize that it is crucial to prioritize the suitability and accuracy of data over mere availability because using readily available but inappropriate (here: for this research unsuitable) data can result in undefined or erroneous outcomes – particularly in the face of edge cases – and should, thus, be avoided [39].

4.2.4.1 Scope of measurement

Operationalization must begin with a clear *scope* of what is to be evaluated. This typically needs to be the end-to-end system; because fairness does not necessarily compose [16], we cannot assume that improving the fairness in some respect for one component of the system will necessarily improve fairness of the system's final output or impact. While it is vital to study different stages and components (e.g., candidate selection [10] or embeddings [47]), they cannot be studied only on their own; downstream impacts are crucial to understanding their contributions to fairness in the system's social impact.

The scope of measurement, therefore, consists of several aspects (some of which are decided in earlier stages, such as problem definition; see Section 4.2.3):

- **What component(s) or intervention(s) are being evaluated?** Some projects will be purely descriptive, seeking to understand the fairness of a current system; others will be incorporated into evaluations of changes proposed for other purposes (e.g., ensuring a model intended to improve user modeling accuracy does not induce unfairness); and still others are to evaluate the effectiveness of a fairness intervention. The scope of measurement needs to be in line with the problem definition (Section 4.2.3) and specified in more (fine-grained) detail.

- **What system aspect(s) are to be evaluated?** As noted above, this usually needs to include fairness of the final system outputs or impacts, but it may also include targeted measurements of other components. For example, an experiment on improving the fairness of candidate selection in a multi-stage research paper recommender system should measure both the fairness of the selected candidates, and the fairness of the final rankings, to assess both (1) if the intervention is behaving as it is intended to (akin to a manipulation check in other research designs) and (2) if it is having the desired effects on the surrounding system.
- **What entity classes are to be considered?** This flows from the selection of stakeholders (see Section 4.4), but operationalization needs to produce a specific metric for users, items, providers, or other entities in the data model; and further, the evaluator must decide whether it is being computed over all entities of that class or a subset of the data. The unit of analysis [44] and aggregation strategy are also important.

4.2.4.2 Inputs to evaluation

At a high level, there are two major computational and data inputs to an evaluation: the system to be evaluated and the data to be used for that evaluation. The system is common to all evaluation types, as is some of the data (consumption or feedback data, content, etc.).

Fairness evaluations often require additional data, particularly for group fairness, where group membership data is required. There is a variety of sources for such data:

- Integrate additional public data sets. For example, [19] combine three external data sources with book consumption data to measure author gender fairness for book recommendations.
- Obtain data from additional sources, such as data markets. Depending on the data source, this may bring significant privacy, ethics, and legal questions.
- Collect or produce data, e.g., by paying for expert data annotations and metadata preparation.
- Use background data available in the specific domain or related domains. Background data, such as demographic information, social indicators, or historical trends, can be a valuable source to fill gaps and enrich the context. Proper validation and alignment with the primary data source are crucial to ensuring that the background data contributes meaningfully.

Great care is needed to appropriately annotate data, particularly for ascribing potentially sensitive identity characteristics to people. For example, the US Program for Cooperative Cataloging has developed recommendations for discerning and recording authors' gender identities [8]. These recommendations disallow inference of gender identity from names or photos, in favor of authors' explicitly-stated identity (preferred) or inferences from pronouns in official biographical material they approved (if the author describes themselves with the pronoun "her", for example, the guidelines allow that as evidence of a female gender identity). Automated inference, while appealing computationally, has significant challenges in terms of its accuracy and fairness as well as ethical and conceptual concerns about its reification of specific ideas of gender and its (dis)respect for autonomy and right to self-identification among the people identified [27, 37]. Each identity has a different set of considerations (which may vary between cultures and regions, for example, in the different ways racial categories function in different countries). However, a similar concern is required for any categorization of people. There are also a range of privacy and regulatory concerns, in some cases prohibiting data collection and in others requiring it [3].

Once the data has been sourced, either internally or externally, operationalization further depends on the nature and encoding of the data. Several key questions about group membership or other fairness-related data attributes affect further design choices, including:

- How complete is the data?
- What biases are in the data? This can be biases in values, biases in errors (e.g., job candidates of particular races are more likely to have erroneous labels), and biases in selection (e.g., label-dependent selection bias [14], where certain label values are more likely to be observed).
- How many and what categories are in the data? E.g., does it only have binary gender, or does it represent non-binary gender identities as well [37]?
- How are entity categories represented? Are they discrete, or does the data represent mixed, partial, or unknown membership?

4.2.4.3 Experiment design

The overall design of the experiment – data splitting, running systems, etc. – for fairness evaluations is not significantly different from other evaluations for accuracy, diversity, novelty, etc., except for the need to incorporate additional data for some fairness constructs. The guidance elsewhere in this report, therefore, applies.

4.2.4.4 Choosing measurements

The actual specific measurements or objectives used to quantify fairness need to align clearly with the problem, the nature of the constructs involved in the problem (e.g., effectiveness or gender), and the practicalities of the data used to compute them.

For example, several metrics for both provider- and consumer-side fairness only operate on discrete binary attributes in which membership is fully known and are therefore difficult or impossible to apply to more realistic settings with multiple groups and unknown or partial membership [39]. This is misaligned with the nature of the construct (many characteristics are not binary), as well as the data practicalities (complete data is extremely rare). Metrics for individual item fairness suffer from other limitations, e.g., they cannot be used to assess systems in isolation but only for relative comparisons across systems [40, 41]

Some of the things that need to be considered for measurement selection include:

- The metric should be a plausible approximation of the problem. This is the most critical consideration because a metric that is computable but does likely not map to the problem likely is not measuring the intended issue.
- For group fairness, the number of groups and the nature of membership [39]. This affects several things, including whether differences or ratios are appropriate, or whether a different way to compare values is needed [23].
- The nature of the impact or resource to be fairly allocated, such as whether it is subtractible (allocation to one person comes at the expense of another) [17, 20]. Zero-sum operationalizations of non-subtractible goods, such as consumer-side utility (one users' good recommendations do not affect another users' bad ones), induce competition where it need not exist [21, 20]. [45] address this for consumer-side equity of utility by using an *positive-sum* metric, the sum of the logs of the total utility for each group, that has optimal reward gain from improving utility for the least-well-served group.
- Metrics should deal in a clear and documented manner with missing data (feedback, group annotations, or other data).

- Metrics and their aggregations should respond well to edge cases such as empty lists, empty groups, etc.
- Whether or not there is a specific target, and if so, what that target is, needs to be clearly specified.
- How fairness should relate to other concerns, such as utility, when appropriate. For example, pursuing equal exposure for items, providers, or groups and exposure proportional to (estimated) utility will yield different metrics [39, 7].

Further, metrics differ in their interpretability and scope of comparability: some can measure fairness in a way that is comparable across data sets or target distributions. The Gini coefficient, for example, is a data-independent measure of resource concentration, and can be used to document that exposure is more heavily concentrated on a smaller set of items in one system or data set than another. On the other hand, expected exposure loss [15] cannot be directly interpreted and can only assess which of several systems better matches the target distribution.

In some cases, it is not necessary to directly measure unfairness, depending on the evaluation goals. Disaggregated evaluation [4, 22] – grouping entities by attribute and computing metric separately for each group – is useful in its own right to assess whether one group is getting greater benefit or harm than another, even without quantifying the difference itself. Distributional evaluation [27] takes this further, looking at distributions across individual entities or within entity groups (e.g., looking at the distribution of utility for consumers of different genders).

4.2.4.5 Iterating on operationalization

Fairness evaluation is not a linear process that can proceed from definition to operationalization to further stages without detours or backtracking, but is often an iterative process. The operationalization needs to be checked against the problem definition to ensure that it accurately captures the construct of interest.

Also, this check should not be done solely by the research team. Following the idea of member checking in qualitative research [9], it is helpful to return to the stakeholders involved in the problem definition to engage them in assessing whether the proposed design captures the concerns they articulated.

4.2.5 Analysis & interpretation

Once the problem is operationalized and the metric results are available, it is important to dedicate substantial time to analyzing and interpreting these results. A core mantra for analyzing results should be: "Think about it!". The results will likely not provide an "obvious" answer to the research question, and we should not assume that an improvement in the metric(s) is enough for a successful experiment. Instead, it is important to get to the meaning of the results and figure out what conclusion the results allow us to make. This is the required basis to figure out how the results can be used to bring this message to the reader (Section 4.2.6).

It has become common practice to perform Exploratory *Data* Analysis (EDA) to define problems and operationalize them to gain deeper insight into the domain and data. Once the results are in, doing Exploratory *Result* Analysis (ERA) is just as important because we need to ensure we understand the results and draw the correct conclusions. We can only form satisfying conclusions to the research problem with a deep analysis.

There is no set-in-stone way of doing analysis. As analysis is an open space, it is also a creative and challenging effort. To provide a starting point, we highlight some questions we could ask ourselves when analyzing results:

- **Do the results "make sense"?** Given the hypothesis or experimental setup, do the results match expectations regarding sign and magnitude? If they do not match expectations, this should be a trigger to take a second look and figure out why they do not match expectations. This could lead to interesting insights, new ideas, or finding bugs in the data or code.
- **How should we interpret the metric(s)?** Is the metric result easily interpretable, or does it require additional effort to understand what a metric value means in the context of this research? Can a particular metric value be interpreted on its own or does it have to be put into relation with others? How can the metric be used to clarify our story?
- **What does the metric measure?** A good practice is to consider what influences a metric to interpret the results better; for instance, what changes in data could lead to positive or negative changes in metric value. Is it possible to cheat the metric so that it improves, though the cause is not favorable? For example, if the difference between two groups in terms of utility is used as a metric, and it should be minimized, then a way to cheat the metric is to reduce utility for the high-performing group and not improve the low-performing group's experience.
- **How do our assumptions impact our results?** Which assumptions was the experiment setup built upon, and how robust are our results to these assumptions? If we changed some of the assumptions, would this change the results? If so, why does it make sense to use the assumptions?

When analyzing, unexpected results will come up. It is valuable to think about these surprises; even if they cannot be explained within the same work, reporting them is encouraged. Reporting such surprising results may lay the ground for future work investigating these phenomena in detail. As a final point, we want to highlight that although the supposed tradeoff between fairness and utility is often claimed, there is not sufficient evidence to conclude that it exists (for details, see [46]). Even if utility metrics may deteriorate slightly, blaming it on a supposed tradeoff is not doing it justice. Further analysis is likely to show how to improve utility without harming fairness so that we can reach systems that are both fair and useful or improve in fairness without a utility loss. As such, it is also valuable for fairness research to report the utility of the system and the impact of the intervention on this utility. Plenty of evidence shows that utility can go up when the system is fairer.

4.2.6 Reporting & sharing

In this section, we highlight some aspects regarding reporting and sharing the scientific work that is particular to fairness in recommender systems. First, it is key to describe and frame the problem addressed in the work clearly, demonstrating why the problem is crucial to address, which may already be a valuable contribution to the community (cf. Section 4.2.3). It is important to note that this is often not about completely solving the fairness problem, but rather about the outcome that is achieved and how it is achieved, e.g., under which assumptions/hypothesis/constraints.

Data sharing: Part of reporting the work involves sharing the data and code used to conduct the research. However, sharing the data in the case of fairness work requires a thorough consideration of the potential harms that may imply and other ethical considerations. For example, it is common to deal with sensitive data about individuals when doing research

on topics with fairness. Therefore, sharing sensitive data should be avoided in such cases, but it may be possible to do so upon request from other researchers if agreeing to non-disclosure of such information. Allowing the work to be reproducible for others while not disseminating sensitive data can be particularly challenging but is critical or better contributing to the community. For example, when working with gender information, releasing such data may harm some individuals. Also, specific annotation errors may occur (e.g., misgendering) that would be harmful to the affected individual if public, while not affecting the statistical results of the work. For such reasons, sharing the annotated data can be particularly undesired by those individuals since it affects them and needs to be done with care and consideration.

Governance: Another consideration involves who will be responsible for the sensitive data collected after the work is published. For example, it is common that a junior researcher is the main person involved in the tasks of creating the required dataset and reporting the results; in such a case, it should be clearly defined who will be the person of contact (who will be in charge of providing this data) if the junior researcher is no longer part of the institution or laboratory. Further, it is important to point out that in some edge cases – that are not common in recommender systems research so far – the best can be not sharing highly sensitive data; for example, if that puts the integrity of some individuals in danger. In such cases, the availability of such data should be taken with utmost care, and it may be appropriate even to delete such data when the research is concluded. Institutional review boards provide guidance in this regard.

Communication: It is crucial to present fairness findings in a manner that is both respectful and objective. For instance, it is more appropriate to describe the observed disparities and then contextualize them within the broader societal or technical challenges than resorting to language that could be perceived as accusatory or judgmental. Adopting a serious and respectful tone fosters a more constructive dialogue. Hence, the report should aim to move the conversation forward, emphasizing that the problem is not entirely solved and highlighting the progress made. It is also important to mention that the previous suggestion applies when writing scientific reports and also when reviewing them. As reviewers, we should not expect that a single work entirely solves a problem; it may be enough to, for example, make a formal definition of the problem that is trying to solve or present a possible solution even if it is not perfect or reaches the maximum score of a given metric. It is essential to recognize that fixing the problem completely is not the only challenge. When defining the problem and proposing a solution, it is important to acknowledge that there may be multiple reasonable choices and ensure that the proposed one aligns with the problem at hand.

Generally, we should avoid making claims that are not supported by evidence and always highlight which specific results are used to draw a specific conclusion. It is crucial to avoid over-claiming as an attempt to demonstrate the value of the work.

Document assumptions: The report should mention the assumptions made when defining the problem. When we define the problem, we always make assumptions, and sometimes, the decisions and hypotheses are taken by a different person, and we need to discover/understand from analyzing the data. Part of operationalization (see Section 4.2.4) involves making these assumptions and understanding others' decisions.

In the report, it is advised to include a section that clearly states the limitations of the work that come from those assumptions. Transparency over the limitations of a work is always desired and should not be used by a reviewer as a way to criticize the work.

Thoughtful and Thorough Limitations: dedicate a section in the paper to clearly state and report the limitations of the work that arise from the underlying assumptions and design choices. A follow-up on the impact or implications of the achieved results helps to emphasize

the potential of the proposed method, increase transparency over the limitations of the work, and open the room for future investigation. Thorough reporting on the limitations of the work should not lead to reviewers underestimating the value of the work. Being explicit about limitations provides avenues for future work and should be seen as a strength.

In summary:

- State clearly that the goal is to move the conversation forward, not to entirely solve the problem.
- Avoid over-claiming your results; clearly state your contributions and their limitations.
- Demonstrate that the problem you are solving is valuable. Avoid solving problems only because the data to solve them is available, and be careful with top problems.
- When sharing data, consider the sensitivity of the dataset and clearly state what decisions you made with regard to the availability of this dataset. With sensitive data, there are more reasons not to share data, even if this harms reproducibility.
- Problem statement: Explain and ground the problem you are helping to solve.
- Explanation and justification: explain how you ended up with your problem definition: argument and justify your choices at every stage.
- Be very clear about assumptions and discuss them in your evaluation.
- Be considerate in the tone of communication: the problems we are tackling deserve a serious and respectful tone and phrasing, and we should avoid being judgmental.
- Do not assume that your choices are the only reasonable ones: for example, the "correct" target does not exist or the "best" algorithm depends on the target.

4.2.7 Conclusion

Since fairness is a complex, nuanced, and context-dependent family of problems, the challenge remains that simple definitions or overly-standardized evaluation approaches are unlikely to be effective. The presented meta-practices shall give guidance on a meta-level. Still, fairness researchers need to thoroughly explore the specific dimension(s) of fairness involved in their targeted research problem and develop a suitable evaluation strategy.

Although we focus on quantitative analysis, this work could also extend to qualitative analysis, particularly in planning and reporting. However, not all the operational aspects discussed for quantitative analysis will be relevant to qualitative analysis.

Additionally, the examples discussed in our work could also be extended to other values, such as environmental considerations. For instance, the principles and methods for evaluating fairness could be adapted to assess recommender systems' sustainability and environmental impact. This adaptation would provide insights into how well these systems align with ecological goals, identify potential tradeoffs, and ensure that environmental considerations are integrated into their operations. Such an approach can help address broader social responsibility issues and ethical impact more comprehensively.

References

1 Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. Beyond personalization: Research directions in multistakeholder recommendation. *arXiv preprint arXiv:1905.01986*, 2019.

2 Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. Discrimination through optimization: How facebook's ad delivery can lead to biased outcomes. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019.

3 Mckane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *FAccT '21*, pages 249–260, New York, NY, USA, March 2021. Association for Computing Machinery.

4 Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, W Duncan Wadsworth, and Hanna Wallach. Designing disaggregated evaluations of AI systems: Choices, considerations, and tradeoffs. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 368–378, New York, NY, USA, July 2021. Association for Computing Machinery.

5 Christine Bauer, Chandni Bagchi, Olusanmi A Hundogan, and Karin van Es. Where are the values? a systematic literature review on news recommender systems. *ACM Transactions on Recommender Systems*, 2(3), 2024.

6 Christine Bauer and Andrés Ferraro. Strategies for mitigating artist gender bias in music recommendation: a simulation study. In *Music Recommender Systems Workshop*, MuRS 2023. Zenodo, 2023.

7 Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. Equity of Attention: Amortizing Individual Fairness in Rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 405–414. ACM, June 2018.

8 Amber Billey, Matthew Haugen, John Hostage, Nancy Sack, and Adam L Schiff. Report of the PCC Ad Hoc Task Group on Gender in Name Authority Records. Technical report, Program for Cooperative Cataloging, October 2016.

9 Linda Birt, Suzanne Scott, Debbie Cavers, Christine Campbell, and Fiona Walter. Member Checking: A Tool to Enhance Trustworthiness or Merely a Nod to Validation? *Qualitative Health Research*, 26(13):1802–1811, November 2016.

10 Amanda Bower, Kristian Lum, Tomo Lazovich, Kyra Yee, and Luca Belli. Random Isn't Always Fair: Candidate Set Imbalance and Exposure Inequality in Recommender Systems. *CoRR*, abs/2209.05000, 2022.

11 Robin Burke. Multisided fairness for recommendation. *CoRR*, abs/1707.00093, 2017.

12 Maarten Buyl and Tijl De Bie. Inherent limitations of AI fairness. *Commun. ACM*, 67(2):48–55, 2024.

13 Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna, editors, *(FAT* '20) Conference on Fairness, Accountability, and Transparency*, pages 525–534. ACM, 2020.

14 Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. Learning under selective labels in the presence of expert consistency. *CoRR*, abs/1807.00905, 2018.

15 Fernando Diaz, Bhaskar Mitra, Michael D Ekstrand, Asia J Biega, and Ben Carterette. Evaluating stochastic rankings with expected exposure. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, CIKM '20. ACM, October 2020.

16 Cynthia Dwork and Christina Ilvento. Fairness under composition. In Avrim Blum, editor, *10th Innovations in Theoretical Computer Science Conference (ITCS 2019)*, volume 124 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 33:1–33:20, Dagstuhl, Germany, 2019. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

17 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.

18 Michael D Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. Fairness in recommender systems. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 679–707. Springer US, New York, NY, 2022.

19 Michael D. Ekstrand and Daniel Kluver. Exploring author gender in book rating and recommendation. *User Modeling and User-Adapted Interaction*, 31(3):377–420, 2021.

20 Michael D Ekstrand and Maria Soledad Pera. Matching consumer fairness objectives & strategies for RecSys. *CoRR*, abs/2209.02662, September 2022.

21 Michael D. Ekstrand, Lex Beattie, Maria Soledad Pera, and Henriette Cramer. Not just algorithms: Strategically addressing consumer impacts in information retrieval. In *Advances in Information Retrieval*, volume 14611 of *Lecture Notes in Computer Science*, pages 314–335. Springer, March 2024.

22 Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the International Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 172–186. PMLR, 2018.

23 Andres Ferraro, Gustavo Ferreira, Fernando Diaz, and Georgina Born. Measuring commonality in recommendation of cultural content to strengthen cultural citizenship. *ACM Transcactions on Recommender Systems*, 2(1), mar 2024.

24 Andrés Ferraro, Xavier Serra, and Christine Bauer. Break the loop: gender imbalance in music recommenders. In *6th ACM SIGIR Conference on Human Information Interaction and Retrieval*, CHIIR '21, pages 249–254, New York, NY, USA, 2021. ACM.

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

26 Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, and Yongfeng Zhang. Towards long-term fairness in recommendation. In Liane Lewin-Eytan, David Carmel, Elad Yom-Tov, Eugene Agichtein, and Evgeniy Gabrilovich, editors, *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining*, pages 445–453. ACM, 2021.

27 Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. Gender Recognition or Gender Reductionism?: The Social Implications of Embedded Gender Recognition Systems. In *CHI '18*, page 8. ACM, April 2018.

28 Hoda Heidari, Vedant Nanda, and Krishna P. Gummadi. On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, volume 97 of *Proceedings of Machine Learning Research*, pages 2692–2701. PMLR, 2019.

29 Karla Hernandez, Stacy L. Smith, Marc Choueiti, and Katherine Pieper. Inclusion in the recording studio?: Gender and race/ethnicity of artists, songwriters & producers across 1,000 popular songs from 2012–2021. Technical report, Annenberg Inclusion Initiative, mar 2022.

30 Maria Heuss, Fatemeh Sarvi, and Maarten de Rijke. Fairness of exposure in light of incomplete exposure estimation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 759–769, New York, NY, USA, 2022. Association for Computing Machinery.

31 Olivier Jeunen and Bart Goethals. Top-K contextual bandits with equity of exposure. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, page 310–320, New York, NY, USA, 2021. Association for Computing Machinery.

32 Michael J. Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, volume 80 of *Proceedings of Machine Learning Research*, pages 2569–2577. PMLR, 2018.

33 Peter Knees and Andrés Ferraro. Bias and feedback loops in music recommendation: Studies on record label impact. In *MORS@ RecSys*, 2022.

34 Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, volume 80 of *Proceedings of Machine Learning Research*, pages 3156–3164. PMLR, 2018.

35 Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI 2019)*, pages 6196–6200. ijcai.org, 2019.

36 Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, 8, November 2020.

37 Christine Pinney, Amifa Raj, Alex Hanna, and Michael D Ekstrand. Much Ado About Gender: Current Practices and Future Recommendations for Appropriate Gender-Aware Information Access. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, pages 269–279. ACM, March 2023.

38 Amifa Raj and Michael D. Ekstrand. Fire dragon and unicorn princess; gender stereotypes and children's products in search engine responses. *CoRR*, abs/2206.13747, 2022.

39 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*, pages 726–736. ACM, July 2022.

40 Theresia Veronika Rampisela, Maria Maistro, Tuukka Ruotsalo, and Christina Lioma. Evaluation measures of individual item fairness for recommender systems: A critical study. *Trans. Recomm. Syst.*, 2023. Just Accepted.

41 Theresia Veronika Rampisela, Tuukka Ruotsalo, Maria Maistro, and Christina Lioma. Can we trust recommender system fairness evaluation? the role of fairness and relevance. In Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, Yi Zhang, Chirag Shah, Craig MacDonald, and Yiqun Liu, editors, *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024)*. ACM, 2024. Just Accepted.

42 Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *FAT* '19*, pages 59–68, New York, NY, USA, January 2019. Association for Computing Machinery.

43 Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 2219–2228, New York, NY, USA, 2018. Association for Computing Machinery.

44 Jessie J. Smith, Lex Beattie, and Henriette Cramer. Scoping fairness objectives and identifying fairness metrics for recommender systems: The practitioners' perspective. In *Proceedings of the ACM Web Conference 2023*, pages 3648–3659, New York, NY, USA, April 2023. Association for Computing Machinery.

45 Lequn Wang and Thorsten Joachims. User fairness, item fairness, and diversity for rankings in two-sided markets. In *Proceedings of the 2021 ACM SIGIR International Conference on*

Theory of Information Retrieval, pages 23–41, New York, NY, USA, July 2021. Association for Computing Machinery.

46 Michael Wick, swetasudha panda, and Jean-Baptiste Tristan. Unlocking Fairness: A Trade-off Revisited. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

47 Le Wu, Lei Chen, Pengyang Shao, Richang Hong, Xiting Wang, and Meng Wang. Learning fair representations for recommendation: A graph-based perspective. In *Proceedings of the Web Conference 2021*, New York, NY, USA, April 2021. ACM.

48 Eva Zangerle and Christine Bauer. Evaluating recommender systems: survey and framework. *ACM Computing Surveys*, 55(8):1–38, 2022.

49 Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking, part I: Score-based ranking. *ACM Computing Survey*, April 2022.

50 Yuying Zhao, Yu Wang, Yunchao Liu, Xueqi Cheng, Charu Aggarwal, and Tyler Derr. Fairness and diversity in recommender systems: a survey. *arXiv preprint arXiv:2307.04644*, 2023.

4.3 Best-Practices for Offline Evaluations of Recommender Systems

Joeran Beel (University of Siegen – Germany, joeran.beel@uni-siegen.de),
Dietmar Jannach (University of Klagenfurt – Austria, dietmar.jannach@aau.at),
Alan Said (University of Gothenburg – Sweden, alansaid@acm.org),
Guy Shani (Ben Gurion University – Beer Sheva – Israel, shanigu@bgu.ac.il),
Tobias Vente (University of Siegen – Germany, tobias.vente@uni-siegen.de),
Lukas Wegmeth (University of Siegen – Germany, lukas.wegmeth@uni-siegen.de)

4.3.1 Introduction

To date, there have been a large number of papers written on challenges and best practices for evaluating recommender systems [6, 9, 13, 17, 18, 36, 38, 24, 36, 48]. Still, papers written and published today often fall short of embracing the practices suggested in prior works. Hence, we aim to suggest practical methods for the recommender systems community to guide researchers toward embracing such practices. We suggest concrete tools that can be immediately implemented in prominent recommendation system research venues such as ACM RecSys and ACM TORS.

We believe that the research community, as a whole, largely agrees on many of the practices that should be embraced. However, it is often the case that individuals are unaware of the many challenges of rigorous evaluation. In addition, adopting these practices often comes at a significant cost in terms of the invested effort and required time. Hence, it may be tempting for researchers not to prioritize such issues when preparing their work for publication.

An example from a methodological perspective based on surveying the literature shows that authors sometimes tune their models on test data, or do not report on how they tuned the hyperparameters of the baselines [38, 41]. Often, we find that certain aspects of the experimental design, e.g., regarding baselines, datasets, or metrics, are not justified beyond the fact others have adopted the same design in previous work. Combined, these aspects may lead to a certain stagnation in our field, as discussed already a decade ago [24, 17, 71]. Similar discussion has been ongoing more recently, e.g., [13, 18, 33].