

De-centering the (Traditional) User: Multistakeholder Evaluation of Recommender Systems

Robin Burke^a, Gediminas Adomavicius^b, Toine Bogers^c, Tommaso Di Noia^d, Dominik Kowald^e, Julia Neidhardt^f, Özlem Özgöbek^g, Maria Soledad Pera^h, Nava Tintarevⁱ, Jürgen Ziegler^j

^a*Department of Information Science, University of Colorado, Boulder, Boulder, Colorado, USA80309,*

^b*Department of Information and Decision Sciences, University of Minnesota, Minneapolis, Minnesota, USA*

^c*IT University of Copenhagen, Copenhagen, Denmark*

^d*Polytechnic University of Bari, Bari, Italy*

^e*Know Center Research GmbH & Graz University of Technology, Graz, Austria*

^f*CD Lab for Recommender Systems, TU Wien, Vienna, Austria*

^g*Norwegian University of Science and Technology, Trondheim, Norway*

^h*TU Delft, Delft, Netherlands*

ⁱ*Maastricht University, Maastricht, Netherlands*

^j*University of Duisburg-Essen, Duisburg, Germany*

Abstract

Multistakeholder recommender systems are those that account for the impacts and preferences of multiple groups of individuals, not just the end users receiving recommendations. Due to their complexity, evaluating these systems cannot be restricted to the overall utility of a single stakeholder, as is often the case of more mainstream recommender system applications. In this article, we focus our discussion on the intricacies of the evaluation of multistakeholder recommender systems. We bring attention to the different aspects involved in the evaluation of multistakeholder recommender systems—from the range of stakeholders involved (including but not limited to producers and consumers) to the values and specific goals of each relevant stakeholder. Additionally, we discuss how to move from

Email addresses: robin.burke@colorado.edu (Robin Burke), gedas@umn.edu (Gediminas Adomavicius), tobo@itu.dk (Toine Bogers), tommaso.dinoia@poliba.it (Tommaso Di Noia), dkowald@know-center.at (Dominik Kowald), julia.neidhardt@tuwien.ac.at (Julia Neidhardt), ozlem.ozgobek@ntnu.no (Özlem Özgöbek), m.s.pera@tudelft.nl (Maria Soledad Pera), n.tintarev@maastrichtuniversity.nl (Nava Tintarev), juergen.ziegler@uni-due.de (Jürgen Ziegler)

theoretical principles to practical implementation, providing specific use case examples. Finally, we outline open research directions for the RecSys community to explore. We aim to provide guidance to researchers and practitioners about how to think about these complex and domain-dependent issues of evaluation in the course of designing, developing, and researching applications with multistakeholder aspects.

Keywords: recommender systems, evaluation, multistakeholder issues

1. Introduction

To develop a holistic view of the operation of a recommender system, it is often important to consider the impact of the system beyond just the primary users who receive recommendations – although the perspectives of such users will always have a primary role in a personalized system. Expanding the frame of evaluation to include other parties, as well as the ecosystem in which the system is deployed, leads us to a multistakeholder view of recommender system evaluation as defined in [2]:

”A **multistakeholder evaluation** is one in which the quality of recommendations is assessed across multiple groups of stakeholders.”

In this article, we provide (i) an overview of the types of recommendation stakeholders that can be considered in conducting such evaluations, (ii) a discussion of the considerations and values that enter into developing measures that capture outcomes of interest for a diversity of stakeholders, (iii) an outline of a methodology for developing and applying multistakeholder evaluation, and (iv) three examples of different multistakeholder scenarios including derivations of evaluation metrics for different stakeholder groups in these different scenarios.

The variety of possible stakeholders we identified that are part of the general recommendation ecosystem is suggested in Figure 1 and defined here, using the terminology from [1, 2]:

Recommendation **consumers** are the traditional recommender system users to whom recommendations are delivered and to which typical forms of recommender system evaluation are oriented.

Item **providers** form the general class of individuals or entities who create or otherwise stand behind the items being recommended.

Upstream stakeholders are those potentially impacted by the recommender system through the provider side of the interaction, but who are not direct contributors of items. For example, in a music streaming recommender, a songwriter may receive royalties based on songs that are played, but it is the musical artist’s performance of the respective song that is the item actually being recommended and listened to.

Downstream stakeholders are those who are impacted by choices that recommendation consumers make, by interacting with chosen items or being impacted by the use or consumption of recommended items. For example, in a recommender system that suggests children’s books to teachers, the children who ultimately get the books (and their parents) are downstream stakeholders from the teachers who are users of the system [15, 18].

System stakeholder is intended to stand in for the organization creating and operating the recommendation platform itself. This group may have a variety of values, including, but not limited to, economic ones that are not necessarily shared by the consumers or providers.

Third-party stakeholders are those individuals or groups who do not have direct interaction with the system that nonetheless have an interest or are impacted by its operation. For example, in a domain such as job recommendation, government agencies charged with ensuring non-discrimination in hiring practices may be considered stakeholders whose requirements are legally binding on the platform operator.

The vast majority of recommender systems research focuses its evaluation only on the perspective of recommendation consumers. However, in most applications, numerous stakeholders are involved in the upstream and downstream parts of the provisioning, recommending, and consumption process. Here, we illustrate this complexity using a (hypothetical) music streaming application as an example—additional examples from other application areas are described in Section 4.

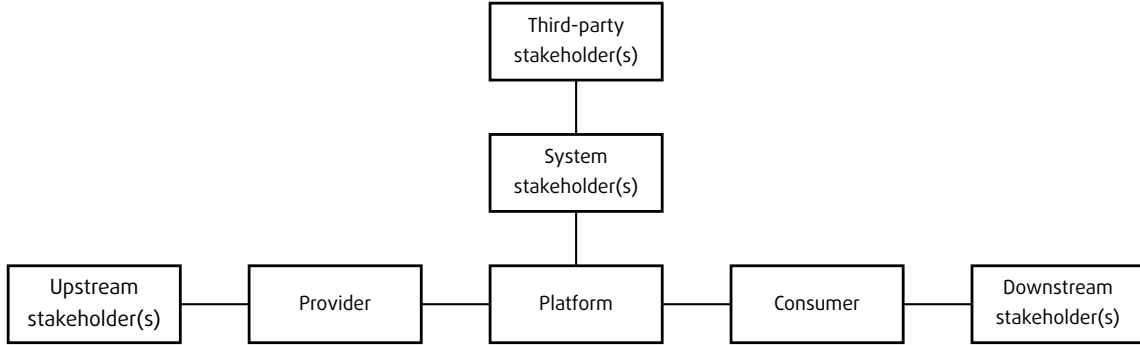


Figure 1: A multistakeholder view of a recommendation ecosystem

Figure 2 shows the different stakeholders involved in the process, with songwriters, artists, and label companies on the content production and provisioning side. The platform (recommender system) plays the role of mediating between upstream and downstream stakeholders. On the downstream side, consumers are the first-line stakeholders, but others may also be affected by the recommendations, e.g., owners of concert venues where recommended artists might appear.

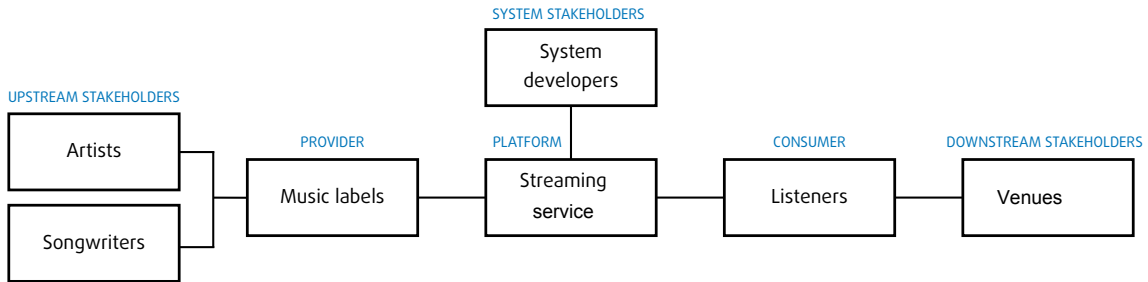


Figure 2: Stakeholder relations for the music streaming example.

Stakeholders pursue specific goals that are driven by values (see Section 2) meant as generic concepts helping an individual to choose the best actions or behaviors [61, 63]. While values are generic concepts and may apply across a wide range of application domains, goals can be seen as intermediate-level objectives that are operationalizations of, for example, a generic human- or business-centric value. Each goal can be assessed by different measures, which may be captured using a variety of concrete measurement methods and

metrics [19]. Unsurprisingly, the goals of different stakeholders may compete with each other, creating the need to balance stakeholder goals in the recommendation process. For the music streaming example, we present sample goals and measures in Table 1. In this context, conflicting goals may be that system operators want to increase the monetary benefit by favoring popular artists and songs which might negatively affect the visibility of long-tail artists who want to build an audience.¹

	Upstream	Provider	System	Consumer	Downstream
Stakeholder	Artist / Songwriter	Music Label	Streaming Service	Listener	Concert Venue
Goals	Monetary reward, Reputation and recognition	Monetary reward, Market development, Product planning	Monetary reward, Customer loyalty	Enjoyment, Wellbeing, Personal development	Monetary reward, Market development, Schedule planning
Measures	Revenue, Royalty, Exposure, User feedback, Playlist inclusion	Revenue, Exposure, Consumption trends, User feedback	Revenue, Customer retention, User feedback	Ratings, Reviews, Music knowledge, Sharing	Ticket & Merchandise Sales, Concertgoer feedback

Table 1: Sample stakeholder goals and measures for the music streaming example.

1.1. Scope

The topic of evaluation touches on many aspects of recommender systems design, implementation and maintenance, more than can be encompassed in a single article. Here, we focus on the problem of principled derivation of evaluation metrics based on an existing system design. We aim to provide guidance to researchers, practitioners, and others who seek to incorporate multistakeholder evaluation into their analysis of recommender system properties and outcomes.

We do not focus on the relationship between evaluation and system design itself, assuming that a system already exists, designed to meet a particular information need, embedded within a particular stakeholder ecosystem. System stakeholders would already have formulated evaluation metrics intended to capture the value that they expect from the recommender system and would have optimized the system to meet those objectives. Sys-

¹We stress that all examples in this discussion are hypothetical and may or may not represent actual stakeholder configurations or goals. For additional perspectives on multi-objective recommendation in music recommendation, see [72].

tem designers may or may not have incorporated diverse stakeholder perspectives in their work, but regardless of the history of design decisions, the impact of the system relative to different stakeholders can still be evaluated.

If developers take a multistakeholder perspective in developing a new system, they would need to engage in many of the analyses outlined in this article to understand how to evaluate the system. Simultaneously, they would have to consider how the recommendation task is defined and how the spectrum of evaluation criteria can be incorporated into the optimization of recommendation models and the delivery of recommendations.

Another topic that we do not address is the design of evaluation methodologies. It is certainly the case that some outcomes (for example, user opinion about the qualities of recommendation lists) can only be measured through surveys or other user studies, whereas other properties (for example, the number of items of a particular type appearing in recommendation lists) can be measured from system outcomes. System properties can be measured in online and offline ways. Readers are referred to the extensive literature on recommender systems evaluation (particularly the overview and surveys in [75, 27, 35]). However, it should be noted that these methodologies are almost exclusively aimed at measuring user-oriented outcomes, with limited research available on evaluation methodologies specifically tailored for other stakeholder outcomes.

1.2. Challenges

Even within the scope that we have chosen for our study, researchers and practitioners face several key challenges that go beyond those typically encountered in recommender systems research and of which they should be aware.

- **Application specificity:** Recommender systems research is, in general, highly domain specific. This specificity is even more pronounced when the larger ecosystem is considered. As our examples make clear (see Section 4), different recommendation applications have different stakeholder configurations and different types of benefits of utility that stakeholders may gain. Even across recommender applications for

which outputs are superficially similar (for example, music playlists), stakeholders may occupy different niches and require different analyses.

- **Access to data:** Typical recommendation datasets have little to no information about non-consumer stakeholders, so it is difficult to understand what are realistic calculations of, for example, revenue distribution among item providers. Additional work will typically be required to gather the data needed to design effective multi-stakeholder evaluations.
- **Context specificity:** Different legal regimes and cultural differences may impose different regulatory requirements on recommender systems, and it is therefore difficult to formulate constraints from third-party stakeholders in a general way.
- **Institutional sensitivity:** There is a strong tradition in research and writing about recommender systems to emphasize the primacy of consumer-side outcomes. This is evident in user interface language of the systems, e.g., via the use of “Recommended for you” and similar labels. Recommendation platforms are often reluctant to publicize or discuss multistakeholder aspects of their systems, even though incorporating such considerations is standard practice.²
- **Adversarial aspects:** Recommendation platforms may actively discourage providers in particular from acquiring knowledge about the platform that might enable strategic activity: for example, misrepresenting their items to gain algorithmic favor. There is no doubt that providers are sometimes incentivized to do this, as the history of search engine spam attests. It is an open research question to design evaluation metrics that can be shared with providers without enabling adversarial behavior.

²As one example, we note that, buried at the bottom of its page on recommendations (<https://www.spotify.com/us/safetyandprivacy/understanding-recommendations>), Spotify states the following “Spotify prioritizes listener satisfaction when recommending content. In some cases, commercial considerations, such as the cost of content or whether we can monetize it, may influence our recommendations.” Such transparency is rare in the industry.

2. Values

Jannach and Zanker [31] mention that, ideally, recommender systems would “*create value in parallel for all involved stakeholders*”. At the same time, it is unavoidable for competing goals to arise, since direct and indirect stakeholders, including the system itself, may have their own perspectives. In this case, to *evaluate* the value created for those involved, we argue that it is imperative to go back to a fundamental and normative question and one that is rarely asked according to Jannach and Zanker [30]: “*What is a good recommendation (in a given context)?*”

To answer this complex question, we posit that one first must look into the values each stakeholder aims for in this multistakeholder process. The concept of ‘value’ has been discussed in the literature from multiple perspectives [28, 70, 2, 9, 69, 25, 53, 69]. Perhaps the most prominent are those referring to the business side of the equation (provider-centered) or the user side (consumer-centered), i.e., the utility of the ultimate consumer. From a more human perspective, values concerning individuals directly or indirectly served by recommender systems and those with societal implications have also been discussed. However, as seen in various practical applications of multistakeholder recommendation tasks, this concept can often be open to multiple interpretations.

In the context of this work, we refer to value as standards or criteria that help an individual to select and evaluate actions or behaviors [61, 63]. With that in mind, for multistakeholder recommender systems, the term value might refer to standards (or even a set of standards) a stakeholder expects or imposes on the recommendation process. The significance of values in system design has been highlighted within the field of human-computer interaction through the development of value-sensitive design processes [50, 23]. In multistakeholder recommender systems, values must be considered when evaluating the ‘goodness’ not just of a recommendation itself, but of the stakeholders that are part of the entire process within the specific contexts and domains in which the systems are deployed.

In the remainder of this section, we review seminal literature that provides background

on the concept of value from different perspectives and its connection to recommender systems. Along the way, we highlight the most common values to reflect on when evaluating multistakeholder recommender systems. It is worth noting that the values we mention are not meant to be an exhaustive list. Instead, they serve as a starting point to encourage reflection among researchers and practitioners to move beyond the more typical ‘producer versus consumer’ perspective and consider a myriad of factors to (simultaneously) account for when evaluating multistakeholder recommender systems.

2.1. Economic and Business-Related Values

When addressing values in the context of multistakeholder recommender system evaluation, economic and business-related values are often considered, especially for providers and system operators.

De Biasio et al. [9] provide a systematic review of value-aware recommender systems, introducing value primarily as an economic concept leading to **monetary reward** (i.e., profit and revenue). They distinguish several aspects that inform the value of monetary reward reflective of a business and economic view, including use value (e.g., increasing revenue by providing useful recommendations), estimated value (related to attractiveness and desirability, such as having a comprehensive music catalog to create recommendations from), cost value (e.g., the economic resources required to distribute a music album to the music streaming platform), and exchange value (the change in value over time, e.g., increase in a music artist’s recognition and popularity on the platform due to effective recommendations).

From this, we observe values related to **user perception** and **customer loyalty**, which are crucial from both a business and economic perspective. These values often relate to “the concepts of quality and personalization, experience and trust, features, and benefits” [9]. For example, in the music industry, a platform that provides highly personalized playlists based on users’ listening history can significantly enhance user satisfaction. This personalization not only helps users discover new music that aligns with their preferences

but also fosters a sense of trust and loyalty towards the platform. Users are more likely to stay subscribed and recommend the service to others if they consistently experience high-quality, relevant recommendations.

The authors in [10] highlight that recommender systems typically serve an organization’s economic values. Besides profit and revenue (i.e., monetary rewards), this might be related to **growth and market development**. For example, music streaming platforms aim to generate profit and attract new users by offering social features like joint playlist creation, which benefit users when their peers are also on the platform. Furthermore, the authors characterize economic recommender systems as systems that exploit “*price and profit information and related concepts from marketing and economics to directly optimize an organization’s profitability.*” Jannach and Adomavicius [28] identify strategic perspectives for both consumers and providers. For consumers, personal utility includes happiness, satisfaction, knowledge, and entertainment. For providers, organizational utility encompasses profit, revenue and growth. In addition, other values, such as **changing user behavior to create demand**, might be relevant. For example, a music streaming platform might recommend emerging artists or newly released tracks to users, encouraging them to explore and adopt new music preferences, thereby creating demand for content that the platform can better monetize.

Jannach and Zanker [31] examine the theory of business models in e-commerce recommender systems and identify the following value-driving aspects: **efficiency** (e.g., the exposure of music artists in recommendation lists or the number of clicks on recommended music tracks), **complementarities** (e.g., creating value through synergies by combining different item types like recommending merchandise articles along with track recommendations of a specific music artist), **lock-in and churn prevention** (e.g., retaining subscribed users by providing meaningful recommendations), and **novelty and product planning** (e.g., finding new fans through recommendations to users who might like an artist’s music or getting inspired to create new music album).

In addition to immediate financial outcomes, recommender systems can enhance a platform’s **brand equity** by creating positive user experiences that increase user satisfaction and loyalty, both critical components of brand equity. Jannach and Jugovac [29] emphasize that well-crafted recommendations enable platforms to differentiate themselves in competitive markets, thereby strengthening their reputation. Similarly, Maslowska et al. [45] suggest that when recommendations align with users’ personal goals, they not only encourage engagement but also create a positive spillover effect on the platform’s brand, enhancing overall trust and loyalty. Recommender systems also open up **cross-selling and up-selling** opportunities in ways that reinforce brand value, such as suggesting premium products or exclusive experiences that align with the brand’s identity (e.g., premium subscriptions or concert tickets). Finally, platforms can also use user interaction data from recommender systems to **develop personalized marketing strategies** while maintaining high privacy standards, a practice that reinforces consumer trust and encourages long-term engagement [45].

2.2. Societal and Human-Centric Values

Beyond economic and business values, societal and human-centric values, which cover other important aspects, are also crucial for businesses and platforms.

Societal and human-centric values for stakeholders in recommender systems focus on ensuring that these systems operate in ways that prioritize humans individually and society as a whole. We find that there are 4 themes of societal and human-centric values for stakeholders in recommender systems that are relevant in the light of evaluation: (i) usefulness, (ii) well-being, (iii) legal and human rights, and (iv) public discourse and safety [69, 70].

Usefulness and enjoyment means that recommendations should meet the needs and expectations of its stakeholders effectively and efficiently [35]. For example, in the case of a music recommender system, users should be able, via the recommender system, to discover new music that they might enjoy and match their tastes. At the same time, usefulness refers to the recommender system’s ability to help music artists get their outputs recommended

to potentially interested listeners. **Control and privacy** is a closely related value that pertains to the degree of influence and customization stakeholders might have over the recommendations that are generated. This includes privacy aspects in a way that users might want to control the amount of their (music) preference data that is shared with the recommender system [52, 69].

Well-being refers to the recommender system’s ability to help its stakeholders to feel satisfied. In the case of a music recommender system, this means that recommendations should influence the experience with the music streaming platform positively, e.g., provide music recommendations to help listeners relax or relieve stress [34]. In this respect, well-being is related to emotional, mental, and physical health. Other related values are **connection, community and social bonding**, e.g., to enable users to connect with like-minded music listeners or to enable music artists to contribute their outputs to a specific community. Thus, also **reputation, recognition and acknowledgement** might be valuable for some stakeholders, e.g., to support music artists in getting their contributions recognized by music listeners [49]. **Personal growth and development** might also be values contributing to well-being in the sense that, e.g., music recommendations could help people explore new music styles and genres, supporting exploration and self-discovery [6].

Concerning legal and human rights, **fairness** might be an important value for stakeholders of a recommender system at evaluation time. For example, the music streaming platform should aim to provide meaningful recommendations to all user groups, independent of, e.g., their musical taste, demographic characteristics, or inclination towards popularity [16, 41, 36, 12]. Additionally, the music recommender system should aim to treat music artists fairly and, in that sense, include novel or (less popular) “niche” artists in the recommendation lists when applicable [37, 67].

Fairness can be related to diversity when the goal is to ensure, for example, that diverse articles and styles are represented in recommendation outputs. Diversity may also have listener-oriented benefits, e.g., help music listeners explore artists that might be new to

them [58, 13]. A recommender system might enable **freedom of expression** as well as **accessibility and inclusiveness** by allowing, e.g., music artists to promote their content independent of the genre or popularity of their music [3, 59]. At the same time, recommender systems should enable users to access the content that they like and enjoy, even when their taste does not match the one of the majority of other music listeners [20, 38]. **Transparency and trustworthiness** might also be an important value for all stakeholders of a recommender system. For instance, music artists might be interested in why they are ranked at a specific position and music listeners might be interested in why a specific artist was recommended to them [65].

Values in the area of public discourse and safety are related to a multitude of societal and human-centric aspects. Here, **societal benefit** goes beyond the satisfaction of individual stakeholders. As an example, a music streaming platform might be interested in fostering cultural enrichment by the recommendation of a diverse set of music [73]. This is related to the value of **tradition and history**, for instance, by recommending local and traditional music, which might be hard to find without the recommender system [21, 42]. The **environmental sustainability** might also be an important value for some recommender systems stakeholders. This may involve implementing energy-efficient recommendation models within the platforms, or promoting local music artists whose concerts offer the opportunity for attendance without requiring extensive travel [46]. Finally, **safety** is concerned with users not being exposed to recommendations of disturbing ethically questionable, or age-inappropriate content. In the case of music recommendations, this could refer to sexist or racist music tracks [47, 56].

2.3. Values in Practice

As we mentioned earlier, the concept of value can be perceived as abstract. Nevertheless, in the context of evaluation of multistakeholder recommender systems, we must be able to somehow quantify it, if the aim is to determine ‘goodness’ for all involved. This task we turn to in Section 3.

3. Methodology

As we previously noted, evaluating recommender systems is a contextually situated problem: different domains, recommendation tasks, and contexts require specific metrics and evaluation setups tailored to that specific recommendation scenario. Multistakeholder evaluation, where the perspectives of other stakeholders are taken into account in addition to that of the consumer, only increases the potential complexity of the evaluation. The complexity of multistakeholder evaluation is demonstrated by the richness and variety of the examples described in Section 4. As a result of this complexity, prescribing exactly which methods to use in which order is impractical. Instead, we attempt to describe best meta-practices for conducting successful multistakeholder evaluation in this section, divided into different stages. We consider this process to be iterative, as findings in a later stage can necessitate returning to an earlier stage, for instance, when learning of a new relevant stakeholder to include or when value shifts occur in one or more stakeholders.

Recall that, in our discussion, we assume that we seek to evaluate an existing recommender system, one that has already been developed to provide a particular recommendation function. Of course, planning for a system’s evaluation should be part of its development and stakeholder consultation should be prioritized in the design and implementation of a multistakeholder recommender system.

3.1. Stakeholders

The cornerstone of multistakeholder evaluation is identifying the relevant stakeholders that will be affected by or affect the recommendation process in some way, as shown in Figure 1. The core parties in any multistakeholder evaluation are the consumers, providers, and the system stakeholders behind the recommendation platform. A sensible first step is to engage with the system stakeholders and gauge their understanding of whom they are recommending to (= consumers) and where the items being recommended come from (= providers). System stakeholders, by virtue of their central role, are also most likely to have the greatest awareness of potential third-party stakeholders whose decisions may

impact the operation of the recommendation platform. Commonly, third-party stakeholders would involve regulatory bodies and institutions; here, the system stakeholder's legal department could help identify relevant regulations (e.g., related to consumer protection) and the right parties to reach out to. Finally, depending on the recommendation scenario, system stakeholders may also be helpful in identifying relevant upstream and downstream stakeholders.

Consumers (or users) have historically played (and continue to play) a central role in recommender systems evaluation. As a result, a common next step would be profiling the consumer stakeholder and the different subgroups this stakeholder category may represent. In addition to interviews with the system stakeholders, any existing market or user research on the user base of the recommendation platform could serve as a valuable foundation for identifying representative subgroups within this user base. A literature review aimed at identifying similar or related recommendation scenarios could also help identify different user groups, especially groups that may be underrepresented in the market research for whatever reason. The system stakeholder should be able to facilitate access to these subgroups, for instance through user research panels, surveys on the website, or customer mailing lists. It is important to recruit a diverse and representative sample of consumers to represent the customer stakeholder and ensure all voices are heard in the evaluation process. Customers should be interviewed or surveyed about what values matter to them in this recommendation scenario (and their relative importance), what goals they have, and how and when they envision using the recommender system. If representative, the principle of saturation could be useful in guiding the sample size required: if additional participants do not reveal any new values, goals, or usage scenarios, then the sample should be representative of the customer stakeholder. Consumers are also a valuable source for identifying possible downstream stakeholders that are worth including in the evaluation process.

The item provider(s) are the general class of individuals or entities who create or oth-

erwise stand behind items being recommended. Historically, they have perhaps been less well represented in recommender systems evaluation, but they play an essential role in multistakeholder evaluation. The number of different individuals or entities that make up the provider stakeholder role may vary greatly between recommendation scenarios: in some cases, only a handful of entities may be providing the items to be recommended, whereas in others they may be as numerous as consumers. Similar to the customer stakeholder, the system stakeholders should be able to facilitate access to the provider stakeholders and help identify which of them carry the biggest weight, without losing sight of the relevant minority providers. Providers are the most valuable source for identifying possible upstream stakeholders that are worth including in the evaluation process. Again, it is important here to recruit a diverse set of representatives for this stakeholder group to ensure that their needs, values, and goals are all met in the evaluation process.

One outcome of interviewing the consumer, provider, and system stakeholders should be the identification of any relevant upstream and downstream stakeholders. This could be supplemented with additional stakeholders identified through a literature review aimed at identifying similar or related recommendation scenarios.

Each stakeholder group should be involved in the process of determining how best to evaluate the quality of recommendations while taking into account the values and goals of each of these stakeholder groups. Qualitative research methods, such as interviews, focus groups, surveys [39], contextual inquiry [60], and co-design [68] could all be beneficial in this process.

3.2. Values and Goals

Once the stakeholders have been identified, the next step involves looking at the values they would want to be part of the recommendation task. Stakeholders' values are at the core of the evaluation process since they drive the modeling of the overall optimization problem. They represent high-level and abstract objectives the stakeholders wish to be satisfied via the use of the recommendation platform [28]. For instance, if the stakeholder

is a music consumer, a possible value is usefulness (of music experience). Conversely, for music providers, a value could be monetary reward or (societal) well-being. It is worth noticing that values may also overlap or partially compete with each other.

The elicitation of values is a fundamental (yet sometimes neglected) step, as it allows the actors involved in designing the system to formulate the goals of each stakeholder involved in a multistakeholder scenario. Going back to the music consumer and provider in our hypothetical example, possible goals might be accuracy and diversity of the recommendation results for the consumer, sell as many items or services as possible, grow the number of users, sell elements over the whole catalog, protect underrepresented groups, or reduce the carbon footprint for the provider. Different from values, goals can be tailored to the specific recommendation domain. A provider may set its goal to grow the number of users listening to classical music, and a consumer may wish to have diverse song recommendation with respect to genre. Goals are more detailed and measurable objectives than values, and they drive the design and implementation of the system through specific evaluation metrics.

3.3. Evaluation Metrics

Formal evaluation metrics provide a way to measure the extent to which the goals of various stakeholders are achieved, i.e., they are measurable proxies towards goals. For example, both consumers and providers are likely to be interested in recommendation accuracy, consumers may be further interested in item discoverability (diversity, novelty, coverage), providers are likely interested in increasing revenue and engagement, and the third-party stakeholders (for instance, regulators) are likely to be interested in consumer-protection-related metrics (representation, fairness, etc.).

Multiple metrics can measure the success of the same goal, depending on the point of view or the aspect we want to highlight. For example, there are different metrics to measure accuracy, e.g., Normalized Discounted Cumulative Gain (NDCG), Mean Reciprocal Rank (MRR), or Recall; we may measure the overall number of items sold in a specific

period or a specific geographical area, the items from the long-tail and the short-head, etc. Depending on the goal, we may have metrics not targeting the overall population of users and stakeholders available in the system.

Some of the specific metrics will naturally come from the prior research literature in recommender systems—the reader may refer to [27] for discussions of some best practices and key metrics in recommender systems evaluation. However, there are clearly opportunities for further metric design, especially for provider-oriented and third-party-oriented stakeholders (i.e., stakeholders that have been under-explored in recommender systems research). Some metrics of these types have been explored in fairness-aware recommendation research [17]. All the metrics must be validated by the target stakeholders (a relevant subset of the overall population is sufficient) to check if they are actually representative of their goals and if they are able to differentiate between high and low utility results. Stakeholders involved in validating the metrics are asked to assess the meaningfulness of the computed results, compared to their goals. A further result of this validation process by the stakeholder can be that of identifying a priority among the metrics. Especially in this phase, a desirable characteristic of a metric is its interpretability and its propensity towards the generation of a human-readable explanation.

As the result of this step, a list of important evaluation metrics (m_1, \dots, m_n) is enumerated, which represents the set of important considerations across multiple stakeholders that need to be taken into account as part of the multistakeholder recommender system evaluation.

3.4. Strategies for Overall Multistakeholder Evaluation

Identifying the list of important evaluation metrics (m_1, \dots, m_n) , as discussed above, provides the ability to evaluate (i.e., to score) a given recommender system R in a multi-dimensional manner. More formally, $\mathbf{S}(R) = (s_1, \dots, s_n)$, where s_i is the performance of R with respect to measure m_i , i.e., $s_i = m_i(R)$. Having multiple evaluation measures raises an important challenge of how to determine the overall (i.e., multistakeholder, multiob-

jective) performance of the system [76]. In particular, given two candidate recommender systems R_A and R_B , where each of which can be evaluated according to the stated list of metrics, $\mathbf{S}(R_A)$ and $\mathbf{S}(R_B)$, how to design a multistakeholder/multiobjective evaluation mechanism \prec_M that allows to determine whether system R_B has superior overall performance to system R_A , i.e., $\mathbf{S}(R_A) \prec_M \mathbf{S}(R_B)$?

Example strategies for developing multistakeholder/multiobjective evaluation mechanisms \prec_M include:

- Weighted (typically linear) aggregation of individual metrics [4, 43] into a single numeric score (as an overall performance), which then allows for a more straightforward comparison of candidate systems.
- Reduction of metric dimensionality by converting some of the individual metrics into constraints [74]. Constraints can be of various types, e.g., hard vs. soft constraints. Hard constraints may indicate the system performance requirements that must be satisfied, which then can be used to filter out candidate systems with inadequate performance. Soft constraints may indicate the relative importance (prioritization) of some metrics, which then can be used to rank the candidate systems accordingly.
- Determining the Pareto frontier of the multidimensional performance vectors of different candidate systems, and measuring the overall performance of a given system as its distance from the Pareto frontier [22]. One key consideration is specifying an appropriate distance metric for multidimensional performance vectors (s_1, \dots, s_n) .
- Learning \prec_M from “ground truth” examples. This could be achieved by providing multiple examples of multidimensional performance vectors $\mathbf{S}(R_i)$ to domain experts, asking them to provide the “ground-truth” judgments regarding the overall performance, and then using machine learning techniques to learn the relationships between the individual metrics and overall performance. For instance, the domain experts could rank pairs of performance vectors at a time, $\mathbf{S}(R_A)$ and $\mathbf{S}(R_B)$, and provide

a ground-truth judgment of whether $\mathbf{S}(R_A) \prec_M \mathbf{S}(R_B)$ or $\mathbf{S}(R_B) \prec_M \mathbf{S}(R_A)$ (or neither, $\mathbf{S}(R_A) \approx_M \mathbf{S}(R_B)$). Learning-to-rank techniques can then be used to build a model for estimating \prec_M from such training data.

More generally, development of multistakeholder/multiobjective evaluation mechanisms \prec_M for recommender systems has connections to several rich research literatures, including multiobjective/multi-criteria optimization [14, 48], multi-criteria decision-making [71] (including its various methodologies, such as data envelopment analysis [7], conjoint analysis [26], multi-attribute utility theory [33]), machine learning [55], and possibly others, which provide promising directions for further research.

Additional considerations impacting the process of overall multistakeholder evaluation include:

- **Stakeholder involvement.** Most of the aforementioned approaches will likely require the involvement of key stakeholders and domain experts, e.g., for determining tradeoffs between individual metrics (leading to decisions regarding relative importance weights for individual metrics or for determining which metrics should be converted to constraints), for obtaining ground-truth judgments about the overall system performance, etc. Therefore, one promising research direction is in the development of *participatory* frameworks [40] that can enable and facilitate stakeholder groups to build algorithmic governance policies for computational decision-making and decision-support systems.
- **Average vs. subgroup vs. individual performance.** It is imperative to establish the perspective for evaluation: Do we evaluate systems in terms of their average performance, or should the distribution of individual performance also be taken into account [57]? For example, does higher average performance also come with much higher individual performance variance (i.e., much worse individual performance for some users/items/etc.), and, if so, what are the right trade-offs? More generally, evaluation at multiple granularities (various subgroup levels) may be of interest.

3.5. *Practical guidelines*

Throughout this section, we have described in detail the best meta-practices for conducting successful multistakeholder evaluation, divided over different stages. We summarize these stages in the list below:

- **Identification of stakeholders.** The inclusion of all relevant stakeholders is essential to the success and representativeness of evaluating a recommender system. Starting with the system stakeholders, consumers and producers, it is important to identify and involve all relevant downstream, upstream and third-party stakeholders. Researchers should consider a range of qualitative research methods and surveys, as well as literature reviews and low-fidelity design processes in doing so.
- **Identification of values and goals.** The goals of the recommender system and expectations for what makes a recommendation good may vary considerably by stakeholder and context. Each stakeholder has different goals and expectations for the recommender system, and these are directly or indirectly tied to the values that matter to them. Qualitative research methods are particularly useful for identifying these values and goals. It is important to keep domain differences in mind when identifying the values and goals. In some domains, certain stakeholders might be affected more seriously by the selection of recommendations provided by the system. It is therefore recommended to undertake a risk analysis at this stage to properly understand the extent of these risks.
- **Selection of evaluation metrics.** Once values and goals have been mapped, they have to be represented by measurable entities, e.g., by selecting existing metrics from prior research that are relevant for a given application context by designing new ones. Part of this process includes determining which metrics to prioritize in the evaluation of the system, as typically multiple metrics can measure the success of the same goal depending on which perspective to highlight. These metrics can vary considerably in

terms of being more qualitative or quantitative in nature, or in terms of representing more short-term or long-term interests. All selected metrics must be validated by the target stakeholders to check whether they are representative of their goals.

- **Strategy selection for overall multistakeholder evaluation.** Choosing a multi-stakeholder approach to recommender systems evaluation may also entail developing a strategy for creating an overall summative evaluation that integrates over the stakeholder perspectives. In other words, evaluating overall performance in a multistakeholder system means that we typically have to deal with a multitude of evaluation metrics, which could also substantially differ between the stakeholder groups. For example, if we know the system performs well for consumer-side metrics, how does this version of the system now work for provider-side metrics? Different strategies for evaluating the overall multistakeholder performance can be employed to find a suitable solution as discussed in Section 3.4.

Also, from the practical perspective, the multistakeholder evaluation methodology—the identification of key stakeholders and their values/goals, the choice of most appropriate individual metrics, the development of specific multistakeholder/multiobjective evaluation mechanisms, and the use of these mechanisms to guide system design and improvement—can be viewed as an iterative process, where researchers and system designers should be aware of all the key steps and can return to iteratively refine any of them.

In reporting on multistakeholder recommendation research, we encourage researchers to include in their discussion the details of stakeholder identification and consultation, the derivation of values and goals, and the justification of metrics in terms of that work. Selbst et al. [64] make the point that formalizations developed in addressing one problem do not necessarily transfer to other contexts. The authors were writing in the context of machine learning fairness, but multistakeholder recommendation is also highly context-specific and similar principles apply.

4. Example Applications and Metrics

Deriving an evaluation metric requires working from a construct, an abstract quality of the recommendation process that we would like to understand, to a concrete proxy of that construct that can be measured and designing a methodology to measure it. The application-specificity of multistakeholder evaluation means that it is difficult to provide such analysis in a general way. With that in mind, we present several specific examples, which serve as means to guide how researchers and industry practitioners might proceed when developing such metrics.

In each of these hypothetical examples, we select a particular stakeholder, as well as a specific value and associated goal, and derive a metric that might be used to evaluate the recommender system relative to that goal. As previously noted, stakeholders are each assumed to have different values, corresponding value-driven goals and potential measures to reach these goals. It is worth reiterating that with these examples, we neither aim to provide a complete set of metrics that one might wish to implement in each of these settings nor highlight the most important metrics. Rather, we seek to illustrate the type of analysis needed to derive such metrics. Moreover, we expect the process of metric selection and development to be iterative rather than linear; this process may even take multiple rounds of consultation and implementation to derive a metric (or set of metrics) that captures a particular stakeholder’s perspective.

The three areas chosen are music streaming, educational resources, and job recommendations. These examples were chosen to highlight different stakeholder perspectives. In music streaming, we focus on musical artists as an example of the provider role. In the recommendation of educational materials, we have a domain where the value of a recommender is more than just consumer taste and yet personalization is still important; here we focus on the student/consumer. Job recommendation is, in many countries, subject to regulation intended to ensure non-discrimination in hiring and is therefore a good place to explore evaluation from the perspective of third-party stakeholders.

4.1. Music Streaming

The first example we consider is streaming music recommendation with the key stakeholders introduced in Figure 2, and also included in Table 1.

In this case, we focus on the providers: the musical artists. There are a variety of values that such individuals might have concerning a distribution platform like a streaming service. We concentrate on the construct of **audience**: an artist will often seek to build a community of individuals who appreciate their particular musical style and contribution (connection, community and social bonding) and might, for example, come to a concert or purchase merchandise (monetary reward) in addition to listening through the streaming service.

A given musical artist might seek to understand to what extent is the recommender system helping them build an audience (use value). One can imagine the system failing in various ways. It might recommend their music to listeners interested in something else, and so the recommendations are not acted upon. Or it might recommend the artist’s music only to listeners who are already fans: helping cement the audience, but not necessarily building it over time. True audience building might only be evident over a long period of time (repeating habitual listening, ticket and merchandise purchases, etc.) so it will probably be necessary to create a short-term proxy for the audience-building potential of a recommender system (growth and market development).

As this is a hypothetical example, our metric is necessarily speculative, but again the aim is to illustrate a process for developing such metrics, not to solve a given evaluation problem. First, we have the problem of measuring an audience from the data available within the streaming service. Let r be the musical artist and let listen count $k_u = \ell(r, u, t)$ be the number of times that user u listens to a track by r over some standard time window t , perhaps one month. The audience A_r can then be defined as the set of individuals for whom this count is greater than some threshold ϵ : $k_u > \epsilon$.

As noted above, measuring audience development can have a long time scale, so a

short term proxy for this quality could be to measure to what extent an artist’s music is being recommended to receptive users. There are multiple ways to determine if a user is receptive³, but for the sake of example, let us assume that we can measure the number n of non-audience listeners (that is, $u \notin A_r$) who were recommended a song by r and then listened to the entire song. Given that musicians have very different numbers of fans, it might make sense to normalize by the size of the artist’s existing audience A_r : $m_r = n/|A_r|$.

As a metric shared with individual providers, a low score on m_r might raise concerns for the artist relative to the recommender system. It would mean that few new listeners are being introduced to their music. For a superstar, this might not be an issue: many people know their music already, but for an emerging artist, it could indicate that the recommender is not working as it should. A higher m_r score does not necessarily mean that their audience is growing, but it does mean that the recommender system is introducing their music to potential new fans. From the system stakeholder point of view, this score could also be aggregated across all providers to understand audience building across the platform’s stable of artists. Its distribution might also be relevant in terms of fairness: are some types of artists better able to build audiences on the platform than others?

4.2. Education

In the context of educational recommender systems, our example focuses on a course content recommender system for secondary school students, possibly integrated within a learning management system (LMS) where the system could track the progress of each student and generate recommendations about what to study next. We illustrate the relationship between value-driven goals and potential measures of each stakeholder, and show how the evaluation perspective changes according to the goal in focus.

In this scenario, teachers provide the content to the recommender system platform both by selecting relevant external content (e.g., educational videos, reference books and

³For example, did the user listen to a second song by the artist, add their songs to a playlist, etc.?

articles) and content generated by themselves. Therefore, we define the external content generators as upstream stakeholders and teachers as provider stakeholders.

The recommender system platform generates course content recommendations for students who are consumer stakeholders and direct users of the system. Parents of the students have an indirect relationship with the generated content (e.g., in the context of recommendation of educational materials for secondary school students, parents might be interested in checking the type of material their children are using) and they are defined as downstream stakeholders. Both upstream and downstream stakeholders have an indirect relationship to the RS platform, which may be relevant to identify and evaluate the value driven goals in a greater picture.

The system stakeholders are responsible for the seamless operation of the recommender system, and they are obliged to ensure that the recommender system platform follows the laws and regulations stated by the school management who is among the *third-party* stakeholders (e.g., the recommended content should be within the corresponding curriculum for each student). Figure 3 illustrates the multistakeholder relations, goals and potential measures in this example scenario.

Based on this example scenario, one point of evaluation of the recommender system platform could be done from the perspective of one of the goals of the consumer stakeholder. More specifically, we could evaluate the recommender system platform from the students' perspective of passing a course, answering the question "How likely is it that a student passes a course when she follows the recommendations from the platform?" (usefulness and enjoyment, as well as personal growth). Although defined from the recommendation consumer's perspective, other stakeholders may benefit the same evaluation. For example, the teacher could use the same measure to understand if the resources she provided to the platform are sufficient in type and quality (usefulness and enjoyment), and the system developers might get an understanding of the relevancy of the recommendations generated by the system beyond click-through rate (use value).

Since the goal of the student is to pass the course at the end of the semester, in this example, we need to evaluate our system at the end of each semester. We assume that the student S_i receives n recommendations every time she uses the system. S_i may choose to accept a recommendation or do another activity on the platform. Therefore, we can measure the number of accepted recommendations by student S_i throughout the semester being n_i . The acceptance of recommendations can be measured in different ways, but for the sake of this example, if the student clicks on any of the recommendations on the list, we assume that the recommendation has been accepted. k_i being the total interaction count of S_i with the system, we can calculate the proportion of the accepted recommendations to the number of whole interactions as $p_i=k_i/n_i$. Finally, at the end of the semester, we calculate the correlation between the student’s final grade in the course and p_i . For the sake of this example, we skip the importance of the order of the recommendations, but an evaluation metric such as NDCG could easily be employed for this purpose. Further, the final metric that correlates the acceptance of recommendations with the student’s final score, could be calculated based on the order of the recommendations, answering the question: Does accepting higher-ranked recommendations from the list correlate with higher student scores?⁴

We should note that the goals of each student may differ or we might be able to identify clusters of students who share the same goals. Therefore, the evaluation methodology could be adjusted according to not only different types of stakeholders but also the differences within one type of stakeholder. This concept of granularity has been discussed in Section 3. Similarly, different stakeholders may have different temporal requirements based on their goals. For example, the students may have a goal for the whole semester (e.g. passing the course), whereas the teachers may have goals that need to be evaluated in a shorter

⁴One might argue for a different indicator of educational value—perhaps the student’s understanding is enhanced in ways less directly measurable—but this equation of final grade with educational value is common in the literature.

term (e.g. understanding if the recommender system platform is helpful for the students to understand the weekly topics).

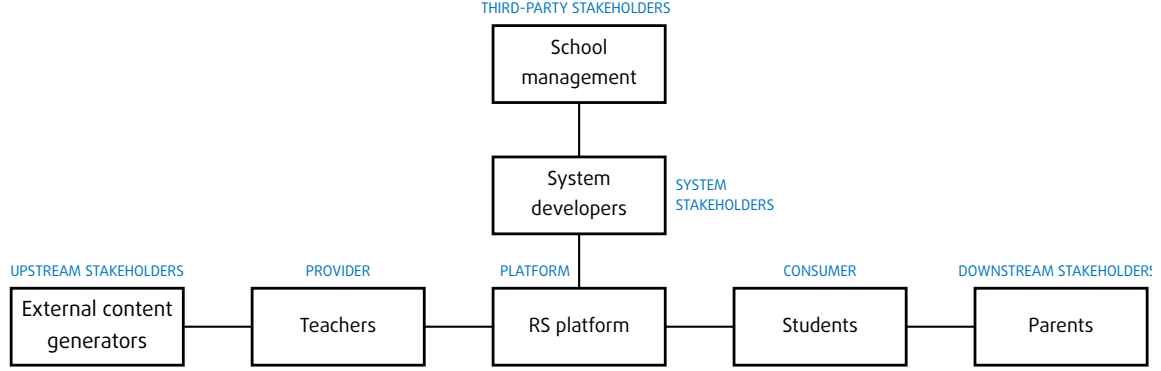


Figure 3: Stakeholder relations for the education example

	Upstream	Provider	System	Third party	Consumer	Downstream
Stakeholder	External content generators	Teachers	RS platform	School management	Students	Parents
Goals	Economic gain, reputation, social benefit	Educating younger generation, social benefit	Economic gain	Social benefit	Passing the course, learning	Educating their children
Measures	Exposure, generating high-quality content	Students learning well, generating high-quality content	Ensuring that the RS works properly, ensuring that the requirements from other stakeholders are satisfied	Ensure that laws and regulations are being followed	Getting good grades, learning the topics well	Reviewing the course material, giving advice to their children

Table 2: Sample stakeholder goals and measures for the education example

4.3. Human Resources

The final example we consider is **candidate recommendation**: recommending suitable candidates for an open job position, also known as talent search. Recruiters often play an important intermediary role in this process by assessing candidates’ qualifications in relation to the job [5]. The candidate identification and assessment process places a great manual burden on recruiters [51] and they would benefit from a system that recommends relevant candidates to supplement their own manual searches. Figure 4 illustrates the different stakeholders involved in this recommendation scenario and is supplemented by Table 3, which displays example goals and measures for each of the stakeholder categories.

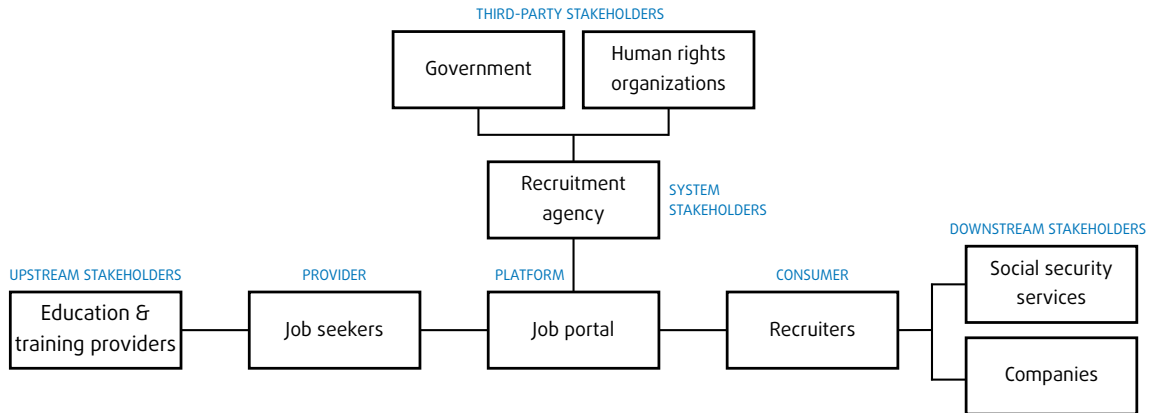


Figure 4: Stakeholder relations for the human resources example

	Upstream	Provider	System	Third party	Consumer	Downstream
Stakeholder	Education & training providers	Job seekers	Job portal	Government	Recruiters	Companies
Goals	Personal development, monetary reward	Personal development, well-being, monetary reward, social bonding	Monetary reward, customer satisfaction, customer loyalty	Employment, social cohesion, economic development, quality of life	Recognition & acknowledgment, personal autonomy, well-being, social bonding	Monetary reward, market development, employee well-being
Measures	Grading scale	Salary increase, working hours	Response rate, % hired, time spent per job, time spent per candidate	Unemployment rate, GDP growth, happiness index	No. of queries issued, time spent per candidate, time spent per job, no. of candidates contacted	Time until position is filled

Table 3: Sample stakeholder goals and measures for the human resources example

This recommendation scenario starts with job seekers by signaling they are open to finding a new job by uploading their CV to the job portal’s CV database, making them the item provider stakeholder. In this scenario, the recruiter is the party receiving the recommendations, making them the consumer stakeholder. The system stakeholder is responsible for creating and operating the candidate recommender system on the job portal, which suggests a slate of relevant candidates to the recruiters. Their values are not necessarily the same as those of the customers and providers. Here we assume that the recruitment agency is the system stakeholder and that they are seeking making their recruiters more efficient through an effective recommender system that allows recruiters to complete job / candidate matches.

Despite paying for the recruitment service, the company with the open job position is not a customer from a multistakeholder evaluation point of view. In this scenario, they instead play the role of downstream stakeholder, as they are impacted by the choices of the recruiters make when assessing, shortlisting and contacting the recommended candidates.

Upstream stakeholders are those potentially impacted by the recommender system, but not direct contributors of items. In the candidate recommendation scenario, education and training providers could function as an upstream stakeholder. These education providers do not have a direct stake in the candidate recommender system but could be interested in learning which skills and competences are most important for a successful matching process, allowing them to update their programs and courses.

Government institutions are an example of third-party stakeholders: they do not have any direct interaction with the job portal, but they have an interest in or are impacted by its operation. A successful candidate recommender system could result in more successful matches between job seekers and companies, affecting important government values such as societal benefit, growth and market development, and well-being.

Government institutions can also have a more direct impact on and interest in the job portal's operation through legislation that ensures non-discrimination in hiring practices, something shared by human rights organizations. Such regulatory practice may impose legally binding requirements on the system stakeholders, affecting the evaluation of the recommended slates of candidates in terms of fairness and protecting underrepresented groups. Job recommendation is therefore a good example to explore evaluation from the perspective of third-party stakeholders.

More specifically, we could evaluate the recommender system platform from the governmental perspective of fairness, answering the question "Given a set of candidates qualified for a job, do the job seekers in both protected and unprotected groups have an equal probability of being contacted?" This question matches the notion of group fairness (or statistical parity), one of the wide variety of fairness metrics [24]. In our scenario, group

fairness is defined as both protected and unprotected groups having an equal probability of being suggested to the recruiter by the recommender system, given they all meet the qualifications set out in the original job posting. Protected groups are defined in terms of sensitive attributes, such as gender, age, ethnicity, and sexual orientation. For example, if a legislative body wanted to ensure gender fairness, an evaluation metric based on group fairness would check whether the difference between the probability of being contacted from the protected group $P(\text{contacted}|\text{qualified} \wedge G = \text{female})$ is equal to the probability of being contacted in the unprotected group $P(\text{contacted}|\text{qualified} \wedge G = \text{male})$ is close to zero.⁵ In an actual multistakeholder evaluation, it would be essential to involve the other stakeholders in determining what fairness means for them, which sensitive attributes are relevant, and how to map this to the most relevant fairness metrics.

5. Concluding Remarks

A holistic understanding of recommender system operation requires considering the perspectives of multiple parties beyond the users receiving recommendations. This area of recommender systems evaluation is relatively underrepresented in the research literature, although, in commercial settings, such considerations have always been an element of recommender system development. Throughout our discussions in Sections 1 to 4, we have emphasized some of the reasons why this work is challenging to conduct and therefore has seen limited research attention.

In our narrative thus far, we have described general properties of multistakeholder recommendation, and methodological approaches to developing relevant metrics, and investigated three hypothetical examples of metric development targeted to different classes of stakeholders. In addressing core ingredients of multistakeholder recommender system evaluation, we hope to inspire reflection of the challenges and possible solutions. In the

⁵Note that assessing whether a given candidate matches the job qualification and to what degree may be complex task in itself.

rest of this section, we discuss some salient challenges that we have identified.

5.1. Transparency / Explainability

Developing multistakeholder metrics and evaluation processes raises the question of to whom such metrics might be reported and made available. Recommender systems evaluation is typically a purely internal matter of engineers or system operators understanding how the recommender operates and seeking to improve it.

However, the types of evaluations that we discuss are different in that they may be of interest to parties who normally have no access to the workings of the recommender system. For example, the musical artists in our streaming example typically have very little insight into how the recommender system treats their content. Earlier, we noted that a given musical artist might seek to understand to what extent is the recommender system helping them build an audience. Such a metric could be shared with artists as a form of explanation to help them understand what the recommender system is doing.

Explanations in a multistakeholder context bring challenges different from explanations targeted only toward consumers. Firstly, different stakeholder groups have different explanatory needs that need to be identified. In the aforementioned example, the artist, and their listeners, have different explanatory needs. The next question is whether one can present different explanations to other stakeholders, or whether the explanation needs to be given to all parties. For example, should we explain item recommendations to individual users and audience building to artists separately? Or is there a single explanation that explains how we resolve the exposure of artists relative to how we weigh user preferences? In other words, if the requirement is to find and generate such a general explanation, then we need systems that can generate a meta-explanation on how tensions were resolved. A particular tension has already been observed, where consumers (depending on the context of recommendation) may prefer less transparency in explanations if it gives better privacy [54, 77]. Members of groups preferred not to disclose sensitive information to other group members, and consumers of advertisements did not want certain information to be used

in explanations (or as a basis of recommendations!). We do not attempt to answer the question of how to generate these kinds of explanations (as this is out of scope), but note that provider-side transparency, let alone generalized explanations, are very little studied in the context of multistakeholder recommendation (evaluation).

5.2. Strategic / Adversarial Considerations

One likely reason that multistakeholder transparency has been little pursued in recommender systems research is the concern that such a facility might be used to enable undesirable adversarial behavior. A web search for the term “YouTube algorithm” yields thousands of hits from search engine optimization (SEO) firms and others advising creators about how to bend the algorithm to their will. Additional information given to providers may enhance their ability to manipulate the algorithm in ways that are not necessarily beneficial to recommendation consumers or the platform. An open research question is how to offer provider-side disclosure in a way that limits adversarial opportunities.

5.3. Governance

Our aim with this article is to help researchers and system designers consider more holistic evaluations of recommender systems, taking multiple stakeholders into account, and examining the impact of the system across stakeholder groups. There is a separate question of governance: who, in the end, has a concrete and effective say in how a recommender system operates? Corporate structures often have a very concrete answer to this question, but as media scholar Nathan Schneider reminds us [62], other models of governance can be and have been applied to online systems. Multistakeholder governance of recommender systems is an interesting question for future research and development.

5.4. Interfaces

Related to the question of governance is the question of interfaces: how do different classes of stakeholders interact with the recommender systems? There is a great deal of study of consumer-side recommendation interfaces, and a wide variety of interface designs

for end users to generate and interact with recommendations. Recommender systems interfaces for other stakeholders do exist, but are rarely the subject of published research. For example, YouTube provides a set of tools within their YouTube Studio application⁶ to enable video creators to see some information about the viewership of their videos, but there are no detailed analytics about how the recommender system is handling their content or ways to interact with the recommender system itself.

The adversarial considerations noted above have no doubt deterred recommender system platforms from offering the kind of transparency into recommender system operations that other stakeholders might find useful. As a result, this is a highly underexplored aspect of multistakeholder recommender systems. Except for a few recent qualitative studies [8, 66], there is relatively little knowledge about provider-side experiences with recommender system interfaces.

5.5. Evaluation Design

The most widely used offline evaluation methodologies in the recommender systems are focused on user-oriented metrics like accuracy. When other stakeholders are considered, for example, in research that studies provider-side outcomes, researchers usually use the same methodology but evaluate the outcomes with provider-oriented metrics. One could imagine alternatives tailored to particular stakeholders: e.g., ensuring that items sampled in the test data set cover all providers, but there is little to no research on such stakeholder-specific evaluations.

Knijnenburg et al. [35] present a well-developed methodology for conducting user studies and interpreting them in terms of user experience. Such metrics might be exactly what is needed to understand different consumer-side aspects of a recommender system. There is no comparable methodology for understanding provider-side experiences of recommendation. It would only make sense to conduct user experience evaluation if an interface for providers

⁶studio.youtube.com

exists, so this research area is downstream from the development of such interfaces.

5.6. Interactive / Conversational Recommendation

As of today, users are accustomed to one-shot static recommendations. Nevertheless, interactive/conversational systems are emerging as a technology that will likely change the nature of user interactions with recommender systems. The final outcome of a conversational session depends on the way the interaction is conducted by both parties: the user (consumer) and the system (that may behave on behalf of the provider). In a multistakeholder scenario, interaction is part of the overall recommendation process, and it is driven by the goals of the two actors involved in the conversation. In fact, depending on the conversation/interaction strategies, the final recommendation can be completely different and push towards the satisfaction of different goals of the involved stakeholders [32]. As a final observation, the interactive process itself may affect the satisfaction of some of the stakeholders' goals. Among others, we may cite the number of interactions to get the final recommendation [11] or the seamless perception of the interactive process [44], but these are solely consumer-side metrics. There is little development of (for example) system-oriented metrics for conversational recommendation.

Acknowledgments

The authors thank the Schloss Dagstuhl – Leibniz Center for Informatics for sponsoring and hosting the Evaluation Perspectives of Recommender Systems: Driving Research and Education seminar, and we also thank Christine Bauer, Alan Said and Evan Zangerle for organizing the seminar.

This publication is supported by the ROBUST project: Trustworthy AI-based Systems for Sustainable Growth with project number KICH3.LTP.20.006, which is (partly) financed by the Dutch Research Council (NWO), RTL, DPG, and the Dutch Ministry of Economic Affairs and Climate Policy (EZK) under the program LTP KIC 2020-2024.

The financial support by the Austrian Federal Ministry of Labour and Economy, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association is gratefully acknowledged.

The work by Toine Bogers was supported by the FairMatch project (Innovation Fund Denmark grant number 3195-00003B). The work of Toine Bogers was also supported by the Pioneer Centre for AI (DNRF grant number P1).

The work by Dominik Kowald was supported by the Austrian FFG COMET – Competence Centers for Excellent Technologies Program and funded by BMK, BMAW, as well as the co-financing provinces Styria, Vienna and Tyrol.

All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

- [1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. Recommender systems as multistakeholder environments. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 347–348, 2017.
- [2] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction*, 30:127–158, 2020.
- [3] Adam Patrick Bell, Atiya Dato, Brent Matterson, Joseph Bahhadi, and Chantelle Ko. Assessing accessibility: an instrumental case study of a community music group. *Music Education Research*, 24(3):350–363, 2022.
- [4] Jürgen Branke, Kalyanmoy Deb, Henning Dierolf, and Matthias Osswald. Finding knees in multi-objective optimization. In Xin Yao, Edmund K. Burke, José Antonio Lozano, Jim Smith, Juan Julián Merelo Guervós, John A. Bullinaria, Jonathan E. Rowe, Peter Tiño, Ata Kabán, and Hans-Paul Schwefel, editors, *Parallel Problem Solving from Nature - PPSN VIII, 8th International Conference, Birmingham, UK, September 18-22, 2004, Proceedings*, volume 3242 of *Lecture Notes in Computer Science*, pages 722–731. Springer, 2004. doi: 10.1007/978-3-540-30217-9_73. URL https://doi.org/10.1007/978-3-540-30217-9_73.
- [5] James A. Breugh. Employee Recruitment: Current Knowledge and Important Areas for Future Research. *Human Resource Management Review*, 18(3):103–118, 2008.
- [6] Òscar Celma and Pedro Cano. From hits to niches? or how popular artists can bias music recommendation and discovery. In *Proceedings of the 2nd KDD workshop on large-scale recommender systems and the netflix prize competition*, pages 1–8, 2008.

- [7] Abraham Charnes, William W. Cooper, Arie Y. Lewin, and Lawrence M. Seiford, editors. *Data Envelopment Analysis Theory, Methodology and Applications*. Springer Science & Business Media, 1995.
- [8] Yoonseo Choi, Eun Jeong Kang, Min Kyung Lee, and Juho Kim. Creator-friendly algorithms: Behaviors, challenges, and design opportunities in algorithmic platforms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–22, 2023.
- [9] Alvise De Biasio, Andrea Montagna, Fabio Aiolli, and Nicolò Navarin. A systematic review of value-aware recommender systems. *Expert Systems with Applications*, page 120131, 2023.
- [10] Alvise De Biasio, Nicolò Navarin, and Dietmar Jannach. Economic recommender systems - a systematic review. *Electronic Commerce Research and Applications*, 63: 101352, 2023.
- [11] Tommaso Di Noia, Francesco Maria Donini, Dietmar Jannach, Fedelucio Narducci, and Claudio Pomo. Conversational recommendation: Theoretical model and complexity analysis. *Inf. Sci.*, 614:325–347, 2022. doi: 10.1016/J.INS.2022.07.169. URL <https://doi.org/10.1016/j.ins.2022.07.169>.
- [12] Karlijn Dinnissen and Christine Bauer. Fairness in music recommender systems: A stakeholder-centered mini review. *Frontiers in big Data*, 5:913608, 2022.
- [13] Tomislav Duricic, Dominik Kowald, Markus Schedl, and Elisabeth Lex. My friends also prefer diverse music: homophily and link prediction with user preferences for mainstream, novelty, and diversity in music. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 447–454, 2021.

- [14] Matthias Ehrgott. *Multicriteria Optimization*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 3540213988.
- [15] Michael D Ekstrand, Ion Madrazo Azpiazu, Katherine Landau Wright, and Maria Soledad Pera. Retrieving and recommending for the classroom. *ComplexRec*, 6 (2018):14, 2018.
- [16] 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 *Conference on fairness, accountability and transparency*, pages 172–186. PMLR, 2018.
- [17] Michael D Ekstrand, Anubrata Das, Robin Burke, Fernando Diaz, et al. Fairness in information access systems. *Foundations and Trends® in Information Retrieval*, 16 (1-2):1–177, 2022.
- [18] Michael D Ekstrand, Maria Soledad Pera, and Katherine Landau Wright. Seeking information with a more knowledgeable other. *Interactions*, 30(1):70–73, 2023.
- [19] Michael D Ekstrand, Lex Beattie, Maria Soledad Pera, and Henriette Cramer. Not just algorithms: Strategically addressing consumer impacts in information retrieval. In *European Conference on Information Retrieval*, pages 314–335. Springer, 2024.
- [20] Andres Ferraro. Music cold-start and long-tail recommendation: bias in deep representations. In *Proceedings of the 13th ACM conference on recommender systems*, pages 586–590, 2019.
- [21] Andres Ferraro, Xavier Serra, and Christine Bauer. What is fair? exploring the artists’ perspective on the fairness of music streaming platforms. In *IFIP conference on human-computer interaction*, pages 562–584. Springer, 2021.

- [22] M. Fleischer. The measure of pareto optima. In Carlos M. Fonseca, Peter J. Fleming, Eckart Zitzler, Kalyanmoy Deb, and Lothar Thiele, editors, *Evolutionary Multi-Criterion Optimization, Second International Conference, EMO 2003, Faro, Portugal, April 8-11, 2003, Proceedings*, volume 2632 of *Lecture Notes in Computer Science*, pages 519–533. Springer, 2003. doi: 10.1007/3-540-36970-8_37. URL https://doi.org/10.1007/3-540-36970-8_37.
- [23] Batya Friedman, Peter H Kahn, Alan Borning, and Alina Huldtgren. Value sensitive design and information systems. *Early engagement and new technologies: Opening up the laboratory*, pages 55–95, 2013.
- [24] Pratyush Garg, John Villasenor, and Virginia Foggo. Fairness metrics: A comparative analysis. In *2020 IEEE international conference on big data (Big Data)*, pages 3662–3666. IEEE, 2020.
- [25] Nada Ghanem, Stephan Leitner, and Dietmar Jannach. Balancing consumer and business value of recommender systems: A simulation-based analysis. *Electronic Commerce Research and Applications*, 55:101195, 2022.
- [26] Paul E. Green and Venkat Srinivasan. Conjoint analysis in marketing: New developments with implications for research and practice. *Journal of Marketing*, 54:3–19, 1990.
- [27] Asela Gunawardana, Guy Shani, and Sivan Yogev. Evaluating recommender systems. In *Recommender Systems Handbook: Third Edition*, pages 547–601. Springer US, 2022.
- [28] Dietmar Jannach and Gediminas Adomavicius. Recommendations with a purpose. In *Proceedings of the 10th ACM conference on recommender systems*, pages 7–10, 2016.
- [29] Dietmar Jannach and Michael Jugovac. Measuring the business value of recommender systems. *ACM Transactions on Management Information Systems (TMIS)*, 10(4): 1–23, 2019.

- [30] Dietmar Jannach and Markus Zanker. Value and impact of recommender systems. In *Recommender systems handbook*, pages 519–546. Springer, 2012.
- [31] Dietmar Jannach and Markus Zanker. Value and impact of recommender systems. *Recommender Systems Handbook*, page 519, 2022.
- [32] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. A survey on conversational recommender systems. *ACM Comput. Surv.*, 54(5):105:1–105:36, 2022. doi: 10.1145/3453154. URL <https://doi.org/10.1145/3453154>.
- [33] Ralph L. Keeney and Howard Raiffa. *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. Cambridge University Press, 1993.
- [34] Peter Knees, Markus Schedl, Bruce Ferwerda, and Audrey Laplante. User awareness in music recommender systems. *Personalized human-computer interaction*, pages 223–252, 2019.
- [35] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. Explaining the user experience of recommender systems. *User modeling and user-adapted interaction*, 22:441–504, 2012.
- [36] Dominik Kowald and Emanuel Lacic. Popularity bias in collaborative filtering-based multimedia recommender systems. In *International Workshop on Algorithmic Bias in Search and Recommendation*, pages 1–11. Springer, 2022.
- [37] Dominik Kowald, Markus Schedl, and Elisabeth Lex. The unfairness of popularity bias in music recommendation: A reproducibility study. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, pages 35–42. Springer, 2020.
- [38] Dominik Kowald, Peter Muellner, Eva Zangerle, Christine Bauer, Markus Schedl, and

- Elisabeth Lex. Support the underground: characteristics of beyond-mainstream music listeners. *EPJ Data Science*, 10(1):14, 2021.
- [39] Mike Kuniavsky. *Observing the user experience: a practitioner's guide to user research*. Elsevier, 2003.
- [40] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. Webuildai: Participatory framework for algorithmic governance. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019. doi: 10.1145/3359283. URL <https://doi.org/10.1145/3359283>.
- [41] Oleg Lesota, Alessandro Melchiorre, Navid Rekabsaz, Stefan Brandl, Dominik Kowald, Elisabeth Lex, and Markus Schedl. Analyzing item popularity bias of music recommender systems: are different genders equally affected? In *Proceedings of the 15th ACM conference on recommender systems*, pages 601–606, 2021.
- [42] Oleg Lesota, Jonas Geiger, Max Walder, Dominik Kowald, and Markus Schedl. Oh, behave! country representation dynamics created by feedback loops in music recommender systems. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 1022–1027, 2024.
- [43] M. Lightner and S. Director. Multiple criterion optimization for the design of electronic circuits. *IEEE Transactions on Circuits and Systems*, 28(3):169–179, 1981. doi: 10.1109/TCS.1981.1084969.
- [44] Ahtsham Manzoor, Wanling Cai, and Dietmar Jannach. Factors influencing the perceived meaningfulness of system responses in conversational recommendation. In Peter Brusilovsky, Marco de Gemmis, Alexander Felfernig, Pasquale Lops, Marco Polignano, Giovanni Semeraro, and Martijn C. Willemsen, editors, *Proceedings of*

the 10th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS 2023) co-located with 17th ACM Conference on Recommender Systems (RecSys 2023), Hybrid Event, Singapore, September 18, 2023, volume 3534 of *CEUR Workshop Proceedings*, pages 19–34. CEUR-WS.org, 2023. URL <https://ceur-ws.org/Vol-3534/paper2.pdf>.

- [45] Ewa Maslowska, Edward C Malthouse, and Linda D Hollebeck. The role of recommender systems in fostering consumers’ long-term platform engagement. *Journal of Service Management*, 33(4/5):721–732, 2022.
- [46] Pavel Merinov. Sustainability-oriented recommender systems. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pages 296–300, 2023.
- [47] Beth A Messner, Art Jipson, Paul J Becker, and Bryan Byers. The hardest hate: A sociological analysis of country hate music. *Popular Music and Society*, 30(4):513–531, 2007.
- [48] Kaisa Miettinen. *Nonlinear Multiobjective Optimization*, volume 12 of *International Series in Operations Research & Management Science*. Kluwer Academic Publishers, Boston, USA, 1998.
- [49] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. Ethical aspects of multi-stakeholder recommendation systems. *The information society*, 37(1):35–45, 2021.
- [50] Jessica K Miller, Batya Friedman, Gavin Jancke, and Brian Gill. Value tensions in design: the value sensitive design, development, and appropriation of a corporation’s groupware system. In *Proceedings of the 2007 ACM International Conference on Supporting Group Work*, pages 281–290, 2007.
- [51] Paolo Montuschi, Valentina Gatteschi, Fabrizio Lamberti, Andrea Sanna, and Claudio

- Demartini. Job recruitment and job seeking processes: How technology can help. *IT Professional*, 16(5):41–49, Sep 2014. ISSN 1941-045X. doi: 10.1109/MITP.2013.62.
- [52] Peter Muellner, Dominik Kowald, and Elisabeth Lex. Robustness of meta matrix factorization against strict privacy constraints. In *43rd European Conference on IR Research, ECIR 2021*, pages 107–119. Springer, 2021.
- [53] Emiliana Murgia, Monica Landoni, Theo Huibers, Jerry Alan Fails, and Maria Soledad Pera. The seven layers of complexity of recommender systems for children in educational contexts. *CEUR Workshop Proceedings*, pages 2449, 5–9, 2019.
- [54] Shabnam Najafian, Geoff Musick, Bart Knijnenburg, and Nava Tintarev. How do people make decisions in disclosing personal information in tourism group recommendations in competitive versus cooperative conditions? *User Modeling and User-Adapted Interaction*, 34(3):549–581, 2024.
- [55] Aviv Navon, Aviv Shamsian, Gal Chechik, and Ethan Fetaya. Learning the pareto front with hypernetworks. *CoRR*, abs/2010.04104, 2020. URL <https://arxiv.org/abs/2010.04104>.
- [56] Council on Communications and Media. Impact of music, music lyrics, and music videos on children and youth. *Pediatrics*, 124(5):1488–1494, 2009.
- [57] Vincenzo Paparella, Vito Walter Anelli, Franco Maria Nardini, Raffaele Perego, and Tommaso Di Noia. Post-hoc selection of pareto-optimal solutions in search and recommendation. In Ingo Frommholz, Frank Hopfgartner, Mark Lee, Michael Oakes, Mounia Lalmas, Min Zhang, and Rodrygo L. T. Santos, editors, *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, pages 2013–2023. ACM, 2023. doi: 10.1145/3583780.3615010. URL <https://doi.org/10.1145/3583780.3615010>.

- [58] Lorenzo Porcaro, Carlos Castillo, and Emilia Gómez Gutiérrez. Diversity by design in music recommender systems. *Transactions of the International Society for Music Information Retrieval*. 2021; 4 (1)., 2021.
- [59] Amrina Ramadhani and Kasiyan Kasiyan. Freedom of expression in music: Controversial song lyrics that challenge social norms. *International Journal of Multicultural and Multireligious Understanding*, 11(1):222–231, 2024.
- [60] Mary Elizabeth Raven and Alicia Flanders. Using contextual inquiry to learn about your audiences. *ACM SIGDOC Asterisk Journal of Computer Documentation*, 20(1): 1–13, 1996.
- [61] Milton Rokeach. The nature of human values. *Free Pres*, 1973.
- [62] Nathan Schneider. *Governable Spaces*. University of California Press, 2024.
- [63] Shalom H Schwartz. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):11, 2012.
- [64] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68, 2019.
- [65] Rashmi Sinha and Kirsten Swearingen. The role of transparency in recommender systems. In *CHI’02 extended abstracts on Human factors in computing systems*, pages 830–831, 2002.
- [66] Jessie J. Smith, Aishwarya Satwani, Robin Burke, and Casey Fiesler. Recommend me? designing fairness metrics with providers. In *2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024*, page to appear, New York, NY, USA, 2024. Association for Computing Machinery.

- [67] Nasim Sonboli, Robin Burke, Michael Ekstrand, and Rishabh Mehrotra. The multi-sided complexity of fairness in recommender systems. *AI magazine*, 43(2):164–176, 2022.
- [68] Marc Steen, Menno Manschot, and Nicole De Koning. Benefits of co-design in service design projects. *International journal of design*, 5(2), 2011.
- [69] Jonathan Stray, Alon Halevy, Parisa Assar, Dylan Hadfield-Menell, Craig Boutilier, Amar Ashar, Chloe Bakalar, Lex Beattie, Michael Ekstrand, Claire Leibowicz, et al. Building human values into recommender systems: An interdisciplinary synthesis. *ACM Transactions on Recommender Systems*, 2(3):1–57, 2024.
- [70] Helma Torkamaan, Mohammad Tahaei, Stefan Buijsman, Ziang Xiao, Daricia Wilkinson, and Bart P Knijnenburg. The role of human-centered ai in user modeling, adaptation, and personalization—models, frameworks, and paradigms. In *A Human-Centered Perspective of Intelligent Personalized Environments and Systems*, pages 43–83. Springer, 2024.
- [71] Evangelos Triantaphyllou. *Multi-Criteria Decision Making Methods: A Comparative Study*. Springer, 2000.
- [72] Moshe Unger, Pan Li, Maxime C Cohen, Brian Brost, and Alexander Tuzhilin. Deep multi-objective multi-stakeholder music recommendation. *NYU Stern School of Business Forthcoming*, 2021.
- [73] Saúl Vargas and Pablo Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 109–116, 2011.
- [74] Yv Haimés Yv, Leon S. Lasdon, and Dang Da. On a bicriterion formulation of the problems of integrated system identification and system optimization. *IEEE*

Transactions on Systems, Man, and Cybernetics, SMC-1(3):296–297, 1971. doi: 10.1109/TSMC.1971.4308298.

- [75] Eva Zangerle and Christine Bauer. Evaluating recommender systems: survey and framework. *ACM Computing Surveys*, 55(8):1–38, 2022.
- [76] Yong Zheng and David (Xuejun) Wang. A survey of recommender systems with multi-objective optimization. *Neurocomputing*, 474:141–153, 2022. doi: 10.1016/J.NEUCOM.2021.11.041. URL <https://doi.org/10.1016/j.neucom.2021.11.041>.
- [77] Dina Zilbershtein, Francesco Barile, Daan Odijk, and Nava Tintarev. Bridging the transparency gap: Exploring multi-stakeholder preferences for targeted advertisement explanations. In *IntRS workshop @ Recsys*, 2024.