Dagstuhl Seminar on Evaluation Perspectives of Recommender Systems: Multistakeholder and Multimethod Evaluation

Robin Burke (Department of Information Science, University of Colorado, Boulder, USA, robin.burke@colorado.edu)
Gediminas Adomavicius (University of Minnesota, Minneapolis, USA, gedas@umn.edu)
Toine Bogers (IT University of Copenhagen, Copenhagen, Denmark, tobo@itu.dk)
Tommaso Di Noia (Polytechnic University of Bari, Bari, Italy)
Dominik Kowald (Know-Center & TU Graz, Graz, Austria, dkowald@know-center.at)
Julia Neidhardt (CDL-RecSys, TU Wien, Vienna, Austria, julia.neidhardt@tuwien.ac.at)
Özlem Özgöbek (Norwegian University of Science and Technology, Trondheim, Norway, ozlem.ozgobek@ntnu.no)
Maria Soledad Pera (TU Delft, Delft, Netherlands, m.s.pera@tudelft.nl)
Jürgen Ziegler (University of Duisburg-Essen, Duisburg, Germany, juergen.ziegler@uni-due.de)

Multistakeholder recommender systems are defined by Abdollahpouri et al. [2] as those that account for "the preferences of multiple parties when generating recommendations, especially when these parties are on different sides of the recommendation interaction." 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 section, we focus our discussion on the intricacies involved in understanding what is the "right" construct required to ensure the proper 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 (beyond producers and consumers) to the values and specific goals of each relevant stakeholder. Additionally, we discuss how to move from theoretical evaluation to practical implementation, providing specific use case examples. Finally, we outline open research directions for the RecSys community to explore. Our aim in this section is to provide guidance to researchers and practitioners about how to think about these complex and domain-dependent issues in the course of designing, developing, and researching applications with multistakeholder aspects.

1 Introduction

To develop a holistic view of a recommender system's operation, 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 be important 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 Abdollahpouri et al. [2]:

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

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

The variety of possible stakeholder orientations is suggested in Fig. 1 and defined here, using the terminology from Abdollahpouri et al. [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 to benefit from items being recommended.
- **Upstream** stakeholders are those potentially impacted by the recommender system but not direct contributors of items. For example, in a music streaming recommender, the songwriter may receive royalties based on streams that are played. Still, it is the musical artist's performance of the song that is the item 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 teachers who use the system [14, 16].
- The **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.
- The **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 an area 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.

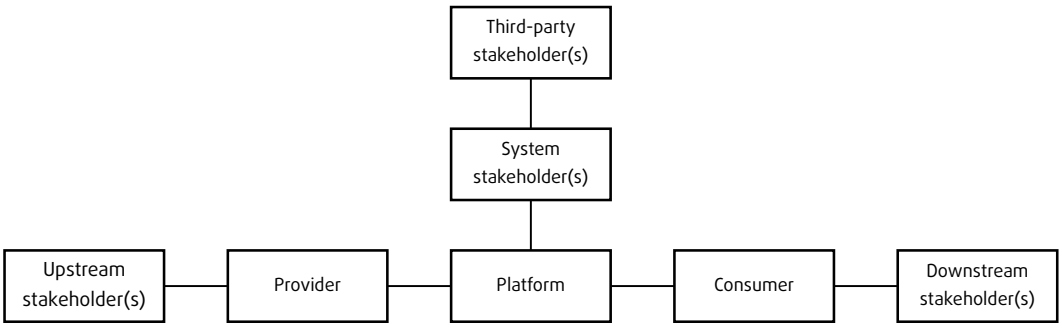

Figure 1 A multistakeholder view of a recommendation ecosystem

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. We illustrate this complexity here with the example of a (hypothetical) music streaming application—additional examples from other application areas are described in Section 4.

Fig. 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 possibly also groups of users may be affected by the recommendations.

Stakeholders pursue specific goals that are driven by values. While values are generic concepts and may apply across a wide range of applications, goals can be considered 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 [17]. Obviously, the goals of different stakeholders may compete with each other, creating the need to balance stakeholder goals in the recommendation process. In the music streaming example, sample goals and measures are given in Table 1. Conflicting goals in this example may be that system operators want to increase monetary benefit by preferring popular artists and songs which might negatively affect the visibility of long-tail artists who want to build an audience[1].

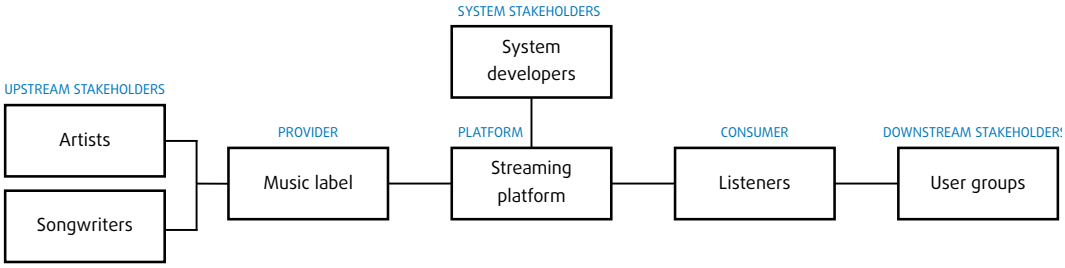

Figure 2 Stakeholder relations for the music streaming example.

	Upstream	Provider	System	Consumer	Downstream
Stakeholder	Artist / Songwriter	Music Label	Streaming Service	Listener	User Groups
Goals	Monetary reward, Reputation and recognition	Monetary reward, Market development, Product planning	Monetary reward, Customer loyalty	Enjoyment, Well-being, Personal development	Enjoyment, Social bonding
Measures	Revenue, Royalty, Exposure, User feedback, Playlist inclusion	Revenue, Exposure, Consumption trends, User feedback	Revenue, Customer retention, User feedback	Ratings, Reviews, Music knowledge, Sharing	Ratings, Reviews, Sharing behavior

Table 1 Sample stakeholder goals and measures for the music example

Multistakeholder evaluation of recommender systems presents additional challenges:

- **Application specificity**: As our examples below make clear, different recommendation applications have different stakeholder configurations and different types of benefits of utility that stakeholders may gain.

[1] We 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 Unger et al. [63].

- **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.
- **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 interface language: "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 especially 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.

2 Values

Jannach and Zanker [27] state 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 [26]: "*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 [25, 61, 2, 9, 60, 22, 44]. 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 the standard (or even set of standards) a stakeholder expects or imposes on the recommendation process. These 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 recommender systems are deployed.

In the rest of this section, we review seminal literature that provides background on the concept of "value" from different perspectives and its connection to recommender

[2] 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.

systems. Along the way, highlight the most common values to consider (in-tandem) *evaluating* multistakeholder recommendation tasks. It is worth noting that the values we mention are not meant as an exhaustive list. Instead, they serve as a starting point to encourage reflection among researchers and (industry) practitioners to move beyond the more typical "producer" versus "consumer" perspectives and consider the 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 systems, 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.

In their work, De Biasio et al. [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 [25] 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 [27] 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).

Beyond these economic and business values, societal and human-centric values, which cover other important aspects, are also crucial for businesses and platforms. These values will be discussed in the following section.

2.2 Societal and Human-centric Values

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 four 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 [60, 61].

Usefulness and enjoyment means that recommendations should meet the needs and expectations of its stakeholders effectively and efficiently [31]. 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 taste. At the same time, usefulness refers to the recommender system's ability to support music artists to 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 their preference data that is shared with the recommender system [60].

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 [30]. 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 people or to enable music artists to contribute their outputs to a specific community. Thus, also **reputation, recognition and acknowledgment** might be valuable for some stakeholders, e.g., to support music artists in getting their contributions being recognized by music listeners [42]. **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** may be an important value for stakeholders of a recommender system at evaluation time. For example, the music stream platform should aim to provide meaningful recommendations to all user groups, independent of, e.g., their musical taste or other demographic characteristics [15, 12]. Additionally, the music recommender system should aim to treat music artists fairly and, in that sense, include novel or "niche" artists in the recommendation lists when applicable [58]. Fairness can be related to **diversity**, which should ensure that recommendations cover a wide set of items to, e.g., help music listeners explore artists that might be new to them [49]. 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 [33, 32, 3, 50]. 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 users [18]. **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 [56].

Furthermore, 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 [64]. 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 [19]. Apart from societal benefits, also the **environmental sustainability** might 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 [39]. 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 [40, 46].

2.3 Values in Practice

As we mentioned earlier, the concept of "value" can be perceived as abstract, and yet, 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.

In Section 3, we offer a theoretical construct to help navigate how to connect values to goals inhered to specific domains and (sub)sets of stakeholders involved, and how these can be operationalized and measured for assessment. Thereafter, in Section 4, we show how we take theory to practice but discuss several examples of multistakeholder recommender system applications.

3 Methodology

As noted elsewhere in this report, 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 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 exact 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 over 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.

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 Fig. 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 be helpful in identifying 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 which values matter to them in this recommendation scenario (and their relative importance), which 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 otherwise stand behind items being recommended. Historically, they have perhaps been less well represented in recommender systems evaluation, but they play an essential role in a multi-stakeholder 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 are 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 of the stakeholder groups 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 [34], contextual inquiry [51], and co-design [59] 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 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 [25]. For instance, if the stakeholder is a music consumer a possible value is *usefulness (of music experience)*. On the other side, 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 step (but 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*, *reduce carbon footprint* for the provider. Differently from values, goals can be tailored to the specific recommendation domain. A provider may set its goal as *grow the number of users listening to classical music*, 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 the metrics.

3.3 Evaluation Metrics

Specific, formal evaluation metrics provide the 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, "long-tailness"), 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., nDCG, MRR, or Recall), we may measure the overall number of items sold in a specific period or in 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 researchers literature in

recommender systems. However, there are clearly opportunities for further metric design, especially so for provider-oriented and third-party-oriented stakeholders (i.e., stakeholders that have been under-explored in recommender systems research). 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 relevant and irrelevant results. Stakeholders validating the metrics are asked to evaluate 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, one 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 $(m_1, \ldots, m_n)$ 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 Multistakeholder Evaluation (Aggregation)

Identifying the list of important evaluation metrics $(m_1, \ldots, m_n)$, as discussed above, provides the ability to evaluate (i.e., to score) a given recommender system R in a multidimensional manner; more formally, $\mathbf{S}(R) = (s_1, \ldots, s_n)$, 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 determine the overall (i.e., multistakeholder, multiobjective) performance of the system [66]. 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, 37] 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 [65]. 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 [20]. One key consideration is specifying an appropriate distance metric for multidimensional performance vectors $(s_1, \ldots, s_n)$.
- 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 research literatures, including multi-objective/multi-criteria optimization [13, 41], multi-criteria decision making [62] (including its various methodologies, such as data envelopment analysis [7], conjoint analysis [23], multi-attribute utility theory [29]), machine learning [45], and possibly others, which provide promising directions for further research.

Additional considerations:

- *Stakeholder involvement.* Most of the above approaches will likely require 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 development of *participatory* frameworks [35] 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.* Important consideration: Do we evaluate systems in terms of their average performance, or should the distribution of individual performance also be taken into account [48]? 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 Use of Multistakeholder Evaluation in System Design and Improvement

Development of evaluation mechanisms $\prec_M$ is important not only for the ability to perform multistakeholder/multiobjective evaluation of recommender systems, but also can also drive decisions for system design and improvement. In particular, the strategies for system design and improvement can be classified as *passive* or *active*.

Passive These are simpler (naive) strategies of using a multistakeholder/multiobjective evaluation mechanism $\prec_M$ to *select* the most advantageous recommender system from a number of (pre-existing) system candidates R_i. These system candidates could possibly be generated even without any multistakeholder considerations in mind (e.g., solely using traditional accuracy-maximizing machine learning approaches) – using $\prec_M$ to select among these candidates would allow incorporating desired multistakeholder considerations to some extent.

Active These are more sophisticated strategies that attempt to *integrate* the multistakeholder/multiobjective evaluation mechanism $\prec_M$ more directly into the system design/optimization process. Two potential sub-categories of active strategies include:

- Adjust/optimize the system recommendations by incorporating $\prec_M$ considerations as a *post-processing* step (e.g., by re-ranking top-N item lists accordingly, etc.), i.e., without directly changing the learning algorithm of the underlying recommender system.

- Adjust/optimize underlying learning algorithms or designing new recommendation algorithms by incorporating $\prec_M$ knowledge directly into the learning process (e.g., by redesigning the loss function accordingly, etc.), so that the produced system recommendations are aligned more directly with the desired multistakeholder considerations.

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. [55] 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, here 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 assumed to each 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.

4.1 Music Streaming

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

We will focus here on the providers, the musical artists. There are a variety of values that such individuals might have with respect to a distribution platform like a streaming service. We concentrate here 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 here 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 timescale, 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 receptive[3], but 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 their music is being introduced 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 interesting 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

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

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 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 a 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). Fig. 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 good or necessary enough (*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. The system generates Top N recommendations for each student. Let's assume that the student S_i receives Top 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 normalized Discounted Cumulative Gain (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 "Is the higher the accepted recommendation on the Top N list, the better the score of the

student?."

We should note that the goals of each student may be different, 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 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 are needed to be evaluated in a shorter term (e.g., understanding if the recommender system platform is helpful for the students to understand the weekly topics).

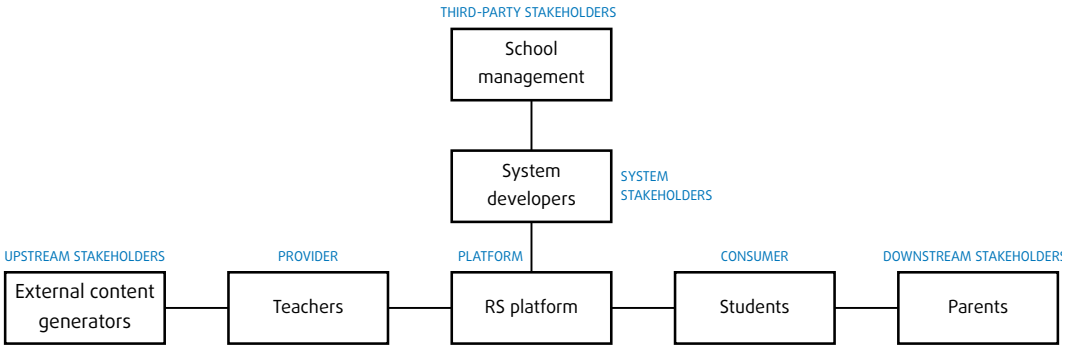

Figure 3 Stakeholder relations for the education example

	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 or estimating person-job fit. Recruiters often play an important intermediary role in this process by assessing candidates' qualifications, such as skills and competences, previous work experience, education level, and remuneration requirements in relation to the job [5]. Much of this candidate identification and assessment process still places a great manual burden on recruiters [43] and a recommender system that suggests relevant candidates to them to approve and supplement with their own manual searches. After shortlisting an acceptable number of candidates, each candidate will be contacted by the recruiter in a (personalized) message, highlighting their match with the job in question and inviting them to apply for the position. Such a recommendation scenario is complex and properly assessing the quality of the candidate recommendations requires

involving multiple stakeholders. Fig. 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.

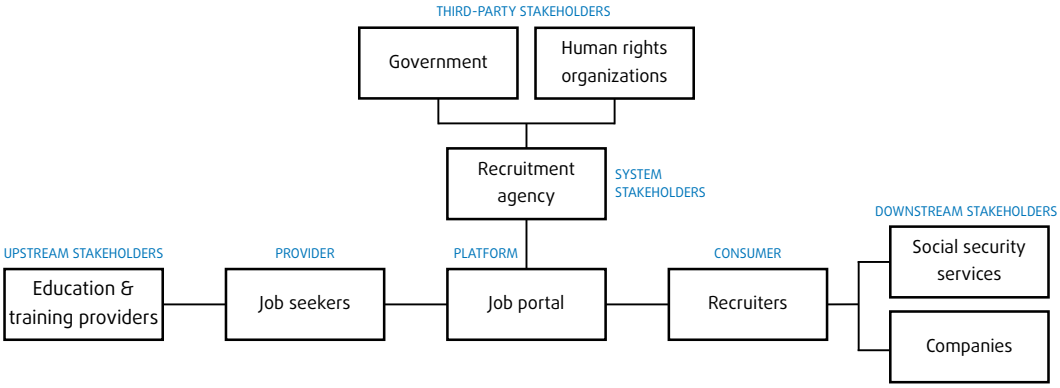

Figure 4 Stakeholder relations for the human resources example

	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

Provider 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. People can be interested in finding a new job for various reasons. Associated values (and potential goals) include (but are not limited to) *personal growth* (e.g., learning new skills and competences or working in new domains), *well-being* (such as a desire to achieve a better work-life balance or working in a job where one's duties have real-world impact), *monetary rewards* (such as a salary increase or better bonus structure), and *connection, community and social bonding* (through friendly colleagues and a supportive working environment). Not all of these goals are equally easy to capture in concrete metrics: a salary increase is easy to measure on paper, but this information is not always accessible to the platform and the system stakeholders. Social bonding is perhaps impractical to capture in a metric.

Consumer The process of recommending candidates to a recruiter starts when a company commissions the recruitment agency that owns the job portal to promote their job posting to relevant candidates. In this scenario, the recruiter is the party receiving the recommendations, making them the **consumer** stakeholder. Like any other employee, recruiters too value their *well-being* and opportunities for *connection, community and social bonding*, but these are affected by the recommendation platform to a lesser degree. Instead, *reputation, recognition and acknowledgment* is more directly related to the recommendation platform, as recruiters

would be interested in seeing their efficiency and effectiveness increase as a result of the recommendations. Efficiency can be measured using many different metrics. In this human-augmented recommendation scenario, the goal is not to replace the human recruiters, but rather support them by reducing the effort they spend on manually searching for candidates. One metric to consider here is the time they spend completing a job, measured from when they first open a new job posting to sending the contact messages to the shortlisted candidates. If the recommender system is able to reduce this total time compared to a scenario without recommendation, the recommender system has likely made them more efficient (barring outside influences or changes to the recruitment process) and has contributed to increased recognition of their work. Other relevant metrics to consider could be the time spent per candidate (which may be more fair to job postings aimed at filling multiple positions), the number of queries issued, or the number of candidates contacted. Another value important to recruiters—albeit one that is hard to capture in metrics—could be *control and privacy*: the introduction of automatic decision support systems and AI-powered tools often induces fears of potential replacement and job loss [24, 36, 47, 52, 53], although research suggests that these fears can be mitigated by additional AI training [24].

System stakeholders 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. In this scenario, the recruitment agency is the system stakeholder, and they are likely to be motivated by *monetary rewards*: making their recruiters more efficient through an effective recommender system would reduce costs per job and allow recruiters to complete more recruiting jobs. The time spent per job or the number of jobs completed per day could be reasonable proxies for this value. Another value could be *customer loyalty*: increasing customer loyalty could be achieved by providing higher-quality matches or providing more matches (which could be at odds with efficiency). Possible metrics for assessing progress towards these goals could be to measure the response rate: if more customers provide a positive response to jobs recommended by a recruiter, this could result in more (high-quality) candidates applying for the position, resulting in greater customer satisfaction and customer loyalty.

Downstream stakeholders 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. Their values are commonly economic in nature, such as *monetary reward* and *growth and market development*. New employees are expected to contribute to the bottom-line of the company. Companies that are currently short-staffed could be seeking to hire new employees to reduce the work pressure on their employees, which flows from the value of employee *well-being*. Such goals could be measured through employee satisfaction surveys, but these are unlikely to be available in the multistakeholder evaluation process. Another potential downstream stakeholder could be social security services: if the recommender system is able to reduce the time spent being unemployed by recommending the right (unemployed) candidate for a job, it could reduce the amount of money that needs to be spent on unemployment benefits. In the end, this benefits society, as this money could be spent on other priorities.

Upstream stakeholders **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. One of their core values is supporting their students' *personal growth*, which is typically measured using a non-binary grading scale. 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.

Third-party stakeholders 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*. These could be quantified using, for instance, the unemployment rate or GDP growth. 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. 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*. Fairness can be measured using a wide variety of metrics [21]. It is therefore essential to involve the other stakeholders in determining what fairness means for them and how to map this to the most relevant fairness metrics.

Human rights organizations are non-governmental organizations that seek to defend the same rights for all members of a society, and represent another third-party stakeholder. In the candidate recommendation scenario, such organizations could be interested in safeguarding values such as *fairness* and *diversity* in the candidate recommendation process, similar to government institutions.

5 Conclusions

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. We discuss above some of the reasons why this work is challenging to conduct and therefore has seen limited research attention.

We have described above general properties of multistakeholder recommendation, and methodological approaches to developing relevant metrics, and investigated three hypothetical examples of metric development. There are many additional aspects of this topic to explore, including:

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 as discussed in this report is typically a purely internal matter of engineers or system operators understanding how the recommender is operating and seeking to improve it. It could be argued that standard summative evaluations of consumer-side outcomes are really only of

interest to the system stakeholder and individual recommendation consumers can assess on their own if the system is working well for them.

The types of evaluations that we discuss here 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 would typically have very little insight into how the recommender system is treating their content. A metric such as the "audience building" one described above could be shared with artists to help them understand what the recommender system is doing. This raises the question of what kinds of transparency the system might want to support relative to such stakeholders. We are not answering this question here, but note that provider-side transparency is very little studied in multistakeholder recommendation.

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 giving advice to creators about how to get the algorithm to bend 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.

5.3 Governance

Our aim in this section 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?[4] Corporate structures often have a very concrete answer to this question, but as media scholar Nathan Schneider reminds us [54], there are other models of governance that 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 application[5] to enable video creators to see some information about the viewership of their videos, but there are no

[4] System governance here is different from data governance as discussed elsewhere in this report.
[5] `https://studio.youtube.com`

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, 57], we know relatively little about provider-side experiences with recommender system interfaces.

5.5 User-centric Evaluation

There is nothing in this discussion that requires metrics are behavioral or off-line. Knijnenburg et al. [31] 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 exists, so this research area is downstream from the development of such interfaces.

5.6 Interactive / Conversational Recommendation

As of today, we are used to one-shot static recommendations. Nevertheless, interactive/-conversational systems are coming to stage possibly changing the way we use recommender systems. The final outcome of a conversational session depends on the way the interaction is conducted from both parties: the user (consumer) and the system (that may behave on behalf of the producer). 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 [28]. As a final observation, the interactive process itself may affect the satisfaction of some 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 [38], but these are solely consumer-side metrics. There is little development of (for example) system-oriented metrics for conversational recommendation.

5.7 Native Multistakeholder Metrics

All the metrics available in the literature so far look at the satisfaction of one single goal per stakeholder. This is the reason why we need aggregation techniques to find the optimal solution to the multistakeholder problem. Unfortunately, aggregation is actually a further approximation of the solution and may need further manual tuning to work properly (see Section 3.4). There could be the need for new metrics which are explicitly conceived to address the multistakeholder problem and then can be configured to satisfy the different goals selected for the problem at hand.

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 Datoo, 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 Tiňo, 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. Breaugh. 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 Matthias Ehrgott. *Multicriteria Optimization*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 3540213988.

14 Michael D Ekstrand, Ion Madrazo Azpiazu, Katherine Landau Wright, and Maria Soledad Pera. Retrieving and recommending for the classroom. *ComplexRec*, 6(2018):14, 2018.

15 Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. All the cool kids, how do they fit

in?: Popularity and demographic biases in recommender evaluation and effectiveness. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the International Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 172–186. PMLR, 2018.

16 Michael D Ekstrand, Maria Soledad Pera, and Katherine Landau Wright. Seeking information with a more knowledgeable other. *Interactions*, 30(1):70–73, 2023.

17 Michael D. Ekstrand, Lex Beattie, Maria Soledad Pera, and Henriette Cramer. Not just algorithms: Strategically addressing consumer impacts in information retrieval. In *Advances in Information Retrieval*, volume 14611 of *Lecture Notes in Computer Science*, pages 314–335. Springer, March 2024. doi: 10.1007/978-3-031-56066-8_25.

18 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.

19 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.

20 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`.

21 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.

22 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.

23 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.

24 Merel Huisman, Erik Ranschaert, William Parker, Domenico Mastrodicasa, Martin Koci, Daniel Pinto de Santos, Francesca Coppola, Sergey Morozov, Marc Zins, Cedric Bohyn, et al. An international survey on ai in radiology in 1,041 radiologists and radiology residents part 1: fear of replacement, knowledge, and attitude. *European radiology*, 31: 7058–7066, 2021.

25 Dietmar Jannach and Gediminas Adomavicius. Recommendations with a purpose. In *Proceedings of the 10th ACM Conference on Recommender Systems - RecSys '16*, page 7–10, Boston, Massachusetts, USA, 2016. ACM Press. ISBN 978-1-4503-4035-9. doi: 10.1145/2959100.2959186. URL `http://dl.acm.org/citation.cfm?doid=2959100.2959186`.

26 Dietmar Jannach and Markus Zanker. Value and impact of recommender systems. In *Recommender systems handbook*, pages 519–546. Springer, 2012.

27 Dietmar Jannach and Markus Zanker. Value and impact of recommender systems. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 519–546. Springer US, New York, NY, 2022. ISBN 978-1-0716-2197-4. doi: 10.1007/978-1-0716-2197-4_14. URL `https://doi.org/10.1007/978-1-0716-2197-4_14`.

28 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`.

29 Ralph L. Keeney and Howard Raiffa. *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. Cambridge University Press, 1993.

30 Peter Knees, Markus Schedl, Bruce Ferwerda, and Audrey Laplante. User awareness in music recommender systems. *Personalized human-computer interaction*, pages 223–252, 2019.

31 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.

32 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.

33 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.

34 Mike Kuniavsky. *Observing the user experience: a practitioner's guide to user research*. Elsevier, 2003.

35 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`.

36 Jian Li and Jin-Song Huang. Dimensions of artificial intelligence anxiety based on the integrated fear acquisition theory. *Technology in Society*, 63:101410, 2020.

37 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.

38 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`.

39 Pavel Merinov. Sustainability-oriented recommender systems. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pages 296–300, 2023.

40 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.

41 Kaisa Miettinen. *Nonlinear Multiobjective Optimization*, volume 12 of *International Series in Operations Research & Management Science*. Kluwer Academic Publishers, Boston, USA, 1998.

42 Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. Ethical aspects of multi-stakeholder recommendation systems. *The information society*, 37(1):35–45, 2021.

43 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.

44 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.

45 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`.

46 Council on Communications and Media. Impact of music, music lyrics, and music videos on children and youth. *Pediatrics*, 124(5):1488–1494, 2009.

47 Olajide Ore and Martin Sposato. Opportunities and risks of artificial intelligence in recruitment and selection. *International Journal of Organizational Analysis*, 30(6): 1771–1782, 2022.

48 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`.

49 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.

50 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.

51 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.

52 Moustaq Karim Khan Rony, Mst Rina Parvin, Md Wahiduzzaman, Mitun Debnath, Shuvashish Das Bala, and Ibne Kayesh. "i wonder if my years of training and expertise will be devalued by machines": Concerns about the replacement of medical professionals by artificial intelligence. *SAGE Open Nursing*, 10:23779608241245220, 2024.

53 Stephan Schlögl, Claudia Postulka, Reinhard Bernsteiner, and Christian Ploder. Artificial intelligence tool penetration in business: Adoption, challenges and fears. In *Knowledge Management in Organizations: 14th International Conference, KMO 2019, Zamora, Spain, July 15–18, 2019, Proceedings 14*, pages 259–270. Springer, 2019.

54 Nathan Schneider. *Governable Spaces*. University of California Press, 2024.

55 Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *FAT* '19*, pages 59–68, New York, NY, USA, January 2019. Association for Computing Machinery. doi: 10.1145/3287560.3287598.

56 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.

57 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.

58 Nasim Sonboli, Robin Burke, Michael Ekstrand, and Rishabh Mehrotra. The multisided complexity of fairness in recommender systems. *AI Magazine*, 43(2):164–176, 2022.

59 Marc Steen, Menno Manschot, and Nicole De Koning. Benefits of co-design in service design projects. *International journal of design*, 5(2), 2011.

60 Jonathan Stray, Alon Halevy, Parisa Assar, Dylan Hadfield-Menell, Craig Boutilier, Amar Ashar, Chloe Bakalar, Lex Beattie, Michael Ekstrand, Claire Leibowicz, Connie Moon Sehat, Sara Johansen, Lianne Kerlin, David Vickrey, Spandana Singh, Sanne Vrijenhoek, Amy Zhang, McKane Andrus, Natali Helberger, Polina Proutskova, Tanushree Mitra, and Nina Vasan. Building human values into recommender systems: An interdisciplinary synthesis. *ACM Transactions on Recommender Systems*, 2(3), jun 2024. doi: 10.1145/3632297.

61 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.

62 Evangelos Triantaphyllou. *Multi-Criteria Decision Making Methods: A Comparative Study*. Springer, New York, NY, USA, 2000. doi: 10.1007/978-1-4757-3157-6.

63 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.

64 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.

65 Yv Haimes 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.

66 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`.