Report from Dagstuhl Seminar 24211

Evaluation Perspectives of Recommender Systems: Driving Research and Education

Christine Bauer[*1], **Alan Said**[*2], **and Eva Zangerle**[*3]

1 **Paris Lodron University Salzburg, AT.** christine.bauer@plus.ac.at
2 **University of Gothenburg, SE.** alansaid@acm.org
3 **University of Innsbruck, AT.** eva.zangerle@uibk.ac.at

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 24211, "Evaluation Perspectives of Recommender Systems: Driving Research and Education", which brought together 41 participants from 16 countries.

The seminar brought together distinguished researchers and practitioners from the recommender systems community, representing a range of expertise and perspectives. The primary objective was to address current challenges and advance the ongoing discourse on the evaluation of recommender systems. The participants' diverse backgrounds and perspectives on evaluation significantly contributed to the discourse on this subject.

The seminar featured eight presentations on current challenges in the evaluation of recommender systems. These presentations sparked the general discussion and facilitated the formation of groups around these topics. As a result, five working groups were established, each focusing on the following areas: theory of evaluation, fairness evaluation, best-practices for offline evaluations of recommender systems, multistakeholder and multimethod evaluation, and evaluating the long-term impact of recommender systems.

Seminar May 20–24 2024 – https://www.dagstuhl.de/24211
2012 ACM Subject Classification Information systems → Recommender systems; Information systems → Evaluation of retrieval results; Human-centered computing → HCI design and evaluation methods
Keywords and phrases Recommender Systems, Evaluation, Information Retrieval, User Interaction, Intelligent Systems
Digital Object Identifier 10.4230/DagRep.14.5.58

1 Executive Summary

Christine Bauer (Paris Lodron University Salzburg, AT, christine.bauer@plus.ac.at)
Alan Said (University of Gothenburg, SE, alansaid@acm.org)
Eva Zangerle (University of Innsbruck, AT, eva.zangerle@uibk.ac.at)

Recommender systems (RS) have become essential tools in everyday life, efficiently helping users discover relevant, useful, and interesting items such as music tracks, movies, or social matches. RS identify the interests and preferences of individual users through explicit input or implicit information inferred from their interactions with the systems and tailor content and recommendations accordingly [13, 16].

* Editor / Organizer

Evaluation Perspectives of Recommender Systems: Driving Research and Education, *Dagstuhl Reports*, Vol. 14, Issue 5, pp. 58–172
Editors: Christine Bauer, Alan Said, and Eva Zangerle
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Evaluation of RS requires attention at every phase of the system life cycle, including design, development, and continuous improvement during operation. High-quality evaluation is crucial for a system's success in practice. This evaluation can focus on the core performance of the system or encompass the entire context in which it is used [3, 7, 8, 10]. Research typically differentiates between system-centric and user-centric evaluation. System-centric evaluation examines algorithmic aspects, such as the predictive accuracy of recommender algorithms. In contrast, user-centric evaluation assesses the user's perspective, including perceived quality and user experience. Comprehensive evaluation must address both aspects since high predictive accuracy does not necessarily meet user expectations [12].

The topic of evaluation, with all its challenges, is currently very relevant and trending. The PERSPECTIVES workshops (organized at ACM RecSys 2021-2023 [14, 15, 11], co-organized by this seminar's organizers) were highly popular and attracted many participants. This interest is further evidenced by the special issue in ACM Transactions on Recommender Systems [1] on evaluation. Recent calls for more impactful RS research [5, 6, 12, 9] highlight that current evaluation practices are too narrow and may not be practically relevant. [4] advocate for more nuanced evaluation methods that meet industry demands. [9] argue that current practices are insufficient as they often overlook side effects or longitudinal impacts. A recent systematic literature study further reveals that current evaluation methods are limited in experiment design, dataset choice, and evaluation metrics [2].

This seminar on evaluation perspectives of RS brought together researchers and practitioners from diverse backgrounds. It aimed to discuss current challenges and advance the ongoing discussion on RS evaluation. The seminar began with eight presentations addressing current challenges in evaluation. These talks initiated the general discussion and helped form groups around these topics. As a result, five working groups were established, each focusing on the following areas:

Working Group 1: Theory of Evaluation

This group focused on the theoretical foundations of RS evaluation. They began by identifying the shortcomings of current evaluation practices and linking these issues to underlying theoretical principles. Key challenges discussed included the selection and configuration of evaluation metrics and the reporting of evaluation results. Section 4.1 outlines the challenges and theoretical perspectives identified in this group.

Working Group 2: Fairness Evaluation

This group focused on exploring paradigms and practices for evaluating the fairness of RS. Given the specific nature of fairness metrics and evaluation requirements for different applications, fairness problems, and goals, the group proposed "best meta-practices", a set of approaches to planning, executing, and communicating rigorous fairness evaluation scenarios. The group's outcome is documented in Section 4.2.

Working Group 3: Best-Practices for Offline Evaluations of Recommender Systems

This working group addressed the topic of offline evaluation, with a specific focus on identifying problems and best practices for this evaluation method. They concentrated on pinpointing the primary challenges related to reproducibility and methodology. Subsequently, they provided guidelines to address these challenges from various perspectives, including those of paper authors, reviewers, editors, and program chairs, as summarized in Section 4.3.

Working Group 4: Multistakeholder and Multimethod Evaluation

This group examined the challenges and complexities in evaluating multistakeholder scenarios, discussing the key aspects that must be considered in such a nuanced environment. Additionally, they explored the transition from theoretical evaluation frameworks to practical implementation. Section 4.4 outlines this work.

Working Group 5: Evaluating the Long-Term Impact of Recommender Systems

This working group concentrated on the long-term perspective and impact of RS and their evaluation. This includes developing suitable long-term measures and conducting social and behavioral research to understand and facilitate aspects such as human behavior, long-term stakeholder goals, and corresponding metrics. Additionally, the group examined practical challenges when evaluating the long-term aspects and impact of RS. This work is presented in Section 4.5.

References

1 Christine Bauer, Alan Said, and Eva Zangerle. Introduction to the special issue on perspectives on recommender systems evaluation. *ACM Transactions on Recommender Systems*, 2 (1), mar 2024. URL `https://doi.org/10.1145/3648398`.

2 Christine Bauer, Eva Zangerle, and Alan Said. Exploring the landscape of recommender systems evaluation: Practices and perspectives. *ACM Transactions on Recommender Systems*, 2(1), mar 2024. URL `https://doi.org/10.1145/3629170`.

3 Joeran Beel, Stefan Langer, Marcel Genzmehr, Bela Gipp, Corinna Breitinger, and Andreas Nürnberger. Research paper recommender system evaluation: a quantitative literature survey. In *Proceedings of the international workshop on reproducibility and replication in recommender systems evaluation*, pages 15–22, 2013.

4 Patrick John Chia, Jacopo Tagliabue, Federico Bianchi, Chloe He, and Brian Ko. Beyond ndcg: behavioral testing of recommender systems with reclist. In *Companion Proceedings of the Web Conference 2022*, pages 99–104, 2022.

5 Dietmar Jannach and Christine Bauer. Escaping the McNamara Fallacy: Towards more impactful recommender systems research. *AI Magazine*, 41(4):79–95, December 2020. ISSN 2371-9621, 0738-4602.

6 Paolo Cremonesi and Dietmar Jannach. Progress in recommender systems research: Crisis? what crisis? *AI Magazine*, 42(3):43–54, 2021.

7 Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, jan 2004. ISSN 1046-8188. `https://doi.org/10.1145/963770.963772`.

8 Dietmar Jannach, Oren Sar Shalom, and Joseph A Konstan. Towards more impactful recommender systems research. In *ImpactRS@ RecSys*, 2019.

9 Gourab K Patro, Lorenzo Porcaro, Laura Mitchell, Qiuyue Zhang, Meike Zehlike, and Nikhil Garg. Fair ranking: a critical review, challenges, and future directions. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 1929–1942, 2022.

10 Alan Said, Domonkos Tikk, Klara Stumpf, Yue Shi, Martha A Larson, and Paolo Cremonesi. Recommender systems evaluation: A 3d benchmark. In *RUE@ RecSys*, pages 21–23, 2012.

11 Alan Said, Eva Zangerle, and Christine Bauer, editors. *Third Workshop: Perspectives on the Evaluation of Recommender Systems (PERSPECTIVES 2023)*, RecSys '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702419. URL `https://doi.org/10.1145/3604915.3608748`.

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

13 Bo Xiao and Izak Benbasat. E-commerce product recommendation agents: Use, characteristics, and impact. *MIS quarterly*, pages 137–209, 2007.

14 Eva Zangerle, Christine Bauer, and Alan Said, editors. *Perspectives on the Evaluation of Recommender Systems (PERSPECTIVES)*, RecSys '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384582. URL `https://doi.org/10.1145/3460231.3470929`.

15 Eva Zangerle, Christine Bauer, and Alan Said, editors. *Second Workshop: Perspectives on the Evaluation of Recommender Systems (PERSPECTIVES 2022)*, RecSys '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392785. URL `https://doi.org/10.1145/3523227.3547408`.

16 Francesco Ricci, Lior Rokach, and Bracha Shapira. *Recommender Systems Handbook*. Springer New York, NY, 3rd edition, 2022.

2 Table of Contents

3 Overview of Talks

3.1 Theory of Evaluation

Neil Hurley (University College Dublin, Ireland, neil.hurley@ucd.ie)

It is commonly believed that empirical evaluations as presented in the recommender system literature are often unclear. The methodology used to carry out the evaluation is not clearly defined, or is incomplete. The justification for this methodology is not articulated. The choice of metrics to compare performance across systems and the configuration of these metrics can seem arbitrary. This should be a major wake-up call to the RS community to sort this out. The theory of evaluation working group will explore metrics, methods, and evaluation protocols for recommender systems performance assessment with a goal of identifying knowledge gaps, where evaluation practices are not backed by sound justifications or a theoretical underpinning. From this exploration, the group will attempt to articulate a way forward for substantially improving the evaluation methodologies that are employed by recommender system developers and are accepted by the community.

3.2 Evaluation in Practice

Bart Goethals (University of Antwerp & Froomle – Belgium, bart.goethals@uantwerpen.be)

Recommender systems are well known to enhance user engagement and generating substantial value for users, providers, and other stakeholders. Online recommender systems are typically evaluated using A/B testing. However, the metrics commonly used for these evaluations, such as click-through rate (CTR), often reflect only short-term user behavior and do not always align with the primary evaluation criterion, which is generally the added value to the provider, such as increased revenue.

Furthermore, recommender systems frequently constitute only a small component of a website. For instance, on an e-commerce site, recommendations may appear in a box labeled *"recommended for you"* on the homepage or below product descriptions on article pages. Consequently, their impact on the overall evaluation criterion can be limited and difficult to quantify.

This presents a significant challenge for recommender system providers in practice: What evaluation methods and metrics should be employed to accurately demonstrate the true value of the recommender system?

3.3 Multistakeholder Evaluation

Robin Burke (Department of Information Science, University of Colorado, Boulder, USA, robin.burke@colorado.edu)

Recommender systems evaluation emphasizes the benefits of recommender systems for end users who receive recommendations and can act on them. An emerging body of research aims to expand the scope of evaluation to consider impacts on a variety of stakeholders beyond these users, typically defined as recommendation consumers. Other stakeholder groups of interest include item providers, those who create or stand behind items that the system recommends, and the organization operating the recommender system, which may have objectives different from those held by either providers or consumers. There is as yet little consensus in the field about appropriate strategies for evaluating the benefit of recommendation to non-consumer stakeholders. What is clear is that, even more than strictly consumer-focused evaluation, there is substantial domain- and application-specificity in how system utility should be defined and evaluated.

3.4 Multi-method Evaluation

Jürgen Ziegler (University of Duisburg-Essen – Duisburg, Germany, juergen.ziegler@uni-due.de)

To obtain a holistic view of a recommender system's quality, applying a single measurement method is not sufficient. Mostly, a combination of different methods will be needed that complement each other depending on the different goals that should be achieved. The motivation for evaluating RS with multiple methods is thus largely driven by the requirement to serve different objectives [1] but also by the needs of different stakeholders affected by the RS [2]. A further purpose of applying multiple methods is to ensure the valid measurement of constructs through cross-validation. Considering the vast space of different methods and metrics available [3], one of the challenges is to select method combinations that provide the most valuable insights into RS quality. While different methods can be characterized along different standard dimensions such as qualitative vs. quantitative measures, or objective versus subjective techniques, combining the perspectives of data-centric and user-centric evaluation appears to provide particularly relevant insights. It has long been shown that data-driven, accuracy-related measures may correlate only weakly with the quality of recommendations as perceived by human users [4]. Combining assessments based on these two perspectives thus is relevant for detecting potential discrepancies between them and for deciding which objectives to prioritize.

However, the application of multiple methods does not imply an overall quality judgment for a particular RS. Determining an overall quality score considering different measures is indeed one of the most difficult challenges in the evaluation of RS. Multi-objective optimization can be a helpful tool for approaching this goal, but weighting the different results and finding acceptable or optimal trade-offs remains an unresolved issue in RS evaluation research. This is particularly true if the goals of different stakeholders need to be taken into account, and

a fair balance between their concerns should be achieved. Importantly, providing more systematic approaches for exploring the trade-off space for RS designs based on multiple methods is a critical, yet under-explored research field. While the final trade-off decisions will need to be taken by the RS provider, ideally in consensus with other stakeholders, the insights gained through different methods can inform and guide the process.

Beyond the application and combination of established methods, there a areas where new methods and metrics will be needed for an effective evaluation. A prominent case are conversational RS which have recently seen a significant boost due to the rapid evolution of NLP techniques, in particular RS based on large foundation models. Considering a broader range of methods including, for example, methods from the fields of linguistics and NLP, seems inevitable to assess the manyfold quality aspects of such systems. Assessing aspects such as dialog strategy, initiative and proactivity in the conversation, or the textual quality of system generated utterances deserve increased attention beyond the mere effectiveness of the recommended items.

References

1 Dietmar Jannach and Himan Abdollahpouri. A survey on multi-objective recommender systems. *Front. Big Data*, 6, March 2023. ISSN 2624-909X. `https://www.frontiersin.org/articles/10.3389/fdata.2023.1157899`. Publisher: Frontiers.

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

4 Sean M. McNee, John Riedl, and Joseph A. Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '06, page 1097–1101, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595932984. URL `https://doi.org/10.1145/1125451.1125659`.

3.5 Evaluation of Fairness

Michael Ekstrand (Drexel University – Philadelphia, PA, USA, mde48@drexel.edu)

Joint work of Michael D. Ekstrand, Anubrata Das, Fernando Diaz, Robin Burke
Main reference Michael D. Ekstrand, Anubrata Das, Robin Burke, Fernando Diaz: "Fairness in Information Access Systems", Found. Trends Inf. Retr., Vol. 16(1-2), pp. 1–177, 2022.
URL http://dx.doi.org/10.1561/1500000079

"Fairness" – ensuring stakeholders of a recommender system are treated fairly in the quantitative and qualitative aspects of their experience [1] – is a complex, multifaceted, contextual, and contested problem that is simultaneously difficult to clearly define and immensely important and impactful for the people affected by a recommender system. Effective fair recommendation work is grounded in specific, well-defined problems that are contextualized in the broader landscape of fairness-related harms.

References

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

3.6 Evaluating the Long-Term Impact of Recommender Systems

Joseph Konstan (University of Minnesota – Minneapolis, USA, konstan@umn.edu)

"Long-term impact" raised the question of how we measure impacts of recommender systems over periods of weeks, months, or longer. Even systems designed around short-term objectives have longer-term effects.

This talk focused on the need for empirical data and longitudinal experiments (in part due to the lack of sufficient theory) and the need to codify best practices).

3.7 Optimizing and evaluating for short- or long-term preferences?

Martijn C. Willemsen (Eindhoven University of Technology & Jheronimus Academy of Data Science)

Recommender Systems are a special case of AI systems as they try to predict user references, and build user models of the user: But for what preferences should we optimize and evaluate? Many recommender systems work optimizes short-term preferences, using behavioral data such as click-streams. But in many cases we might like to extend that approach and take a more forward looking perspective, predicting long-term, aspirational preferences, for example to live a more healthy live or get a more diverse taste of music. How should we evaluate our systems for such long-term preferences and how is that different from short-term preferences?

3.8 Proposal for Evidence-based Best-Practices for Recommender Systems Evaluation

Joeran Beel (University of Siegen / Recommender-Systems.com – Siegen, Germany, joeran.beel@uni-siegen.de)

[1] I recall vividly when more than a decade ago – I was a PhD student – *Konstan & Adomavicius* warned that *"the recommender systems research community [...] is facing a crisis where a significant number of research papers lack the rigor and evaluation to be properly judged and, therefore, have little to contribute to collective knowledge* [24]". Similar concerns were already voiced two years earlier by [17]. Over the following years, many more researchers expressed criticism of the evaluation practices in the community [19, 38, 36, 10], myself included [8, 6, 8, 37]. The situation may have somewhat improved in the past years due to more awareness in the community [19], the reproducibility track at the ACM RecSys conference, innovative submission formats like "result-blind reviews" [7] via registered reports

[1] Please note that I used ChatGPT to improve my writing. I wrote all the sentences first myself and then asked ChatGPT for each paragraph to improve the writing but keep the structure.

at ACM TORS, and several new software libraries, including Elliot [1], RecPack [27], Recbole [45], and LensKit-Auto [42]. Yet the decade-old criticism by *Konstan & Adomavicius* remains as true today as it was a decade ago.

Konstan & Adomavicius proposed that, among others, best-practice guidelines on recommender systems research and evaluations might offer a solution to the crisis [24]. In their paper, they also presented results from a small survey that indicated that such guidelines would be welcomed by many members of the community. However, to my knowledge, no comprehensive guidelines or checklists have been specifically created for the recommender systems community, or at least they have not been widely adopted. Recently, I attempted to develop guidelines for releasing recommender systems research code [4], based on the NeurIPS and 'Papers with Code' guidelines [44], but progress has been limited.

I echo the demand by *Konstan & Adomavicius* [24] for the recommender systems community to establish best-practice guidelines and/or checklists for researchers and reviewers. Such guidelines would facilitate the conduct of "good" research, and they would assist reviewers in conducting through reviews. By "good research" I primarily mean reproducible research with a sound methodology. But "good' research also refers to research that others easily can build upon, e.g. because data and code are available; research that is ethical; and research that is sustainable, e.g. because no resources were wasted.

My vision is best-practice guidelines that are not merely a collection of opinions but are instead grounded in empirical evidence. This approach would be analogous to the medical field, where guidelines for practitioners are justified based on empirical research findings. Additionally, these medical guidelines indicate the degree of consensus among experts, allowing medical practitioners to understand how widely accepted each best practice is. In areas with less expert consensus, deviations from the best practice by practitioners would be more acceptable. This model ensures that guidelines are both scientifically robust and flexible.

In my view, best-practice guidelines for recommender systems research and evaluation should include the following components in addition to the best practices themselves:

1. Justification: A justification for the best practice, ideally based on empirical evidence.
2. Confidence: An estimate of how sound the evidence is.
3. Severity: An estimate of the importance of the best practice and the potential consequences of not following it.
4. Consensus: The degree of agreement within the community or among experts that the proposed best practice is indeed a best practice.

Table 1 illustrates what a best practice may look like, using the example of random seeds. A random seed is an initial value for a pseudo-random number generator, ensuring that the sequence of random numbers it produces is reproducible. This reproducibility is crucial for consistent experiment results, fair comparisons between different algorithms, and reliable debugging. For instance, when splitting a dataset into training and testing sets, using a fixed random seed ensures the same split is produced each time. This consistency allows researchers to compare the performance of different algorithms on identical data splits, ensuring that any performance differences are due to the algorithms themselves and not variations in the data splits. Generating random random-seeds is not a trivial task, and dedicated tools exist for it [16].

Creating a preliminary set of guidelines for recommender systems evaluation should be straightforward. Existing communities, particularly in machine learning, already have robust best-practice guidelines and checklists. Notably, NeurIPS [28, 31] and the AutoML conference [3] offer guidelines that could be adapted for recommender system experiments with relatively

Table 1 Best Practices for Random Seeds (Example).

Random Seeds Best-Practice	1) Experiments must be repeated (n>=5) with different random seeds each time. This is true for each aspect of an experiment that requires randomness. This includes splitting data and initializing weights in neural networks. 2) The exact random seeds used for experiments must be reported in the paper or the code.
Justification	[43] showed that when random seeds differed – i.e. data splits contained different data due to randomness – the performance of the same algorithm, with the same hyper-parameters on the same dataset(s) varied by up to 12% [43]. In contrast, repeating and averaging experiments with different random seeds, led to a maximum difference of only around 4%. This means, if only a single run had been conducted, the results could be up to 6% above or under the 'true' result, possibly more. By repeating the experiments, the difference would have been only ±2% in the worst case. The variance depended on the applied metrics, cut-offs, datasets, and splitting methods (lower variance for cross-fold validation, higher variance for hold-out validation). Therefore, repeating experiments with different random seeds ensures that the reported result is closer to the 'true' result. Reporting the exact random seeds is also a prerequisite (besides many other factors) for an exact replication of experiments. A researcher who wants to replicate an experiment and who uses the identical random seeds as the original researcher, will have the same data in the train and validation splits as the original researcher. Knowing the exact random seeds also makes it easier to detect fraudulent behavior such as cherry picking.
Severity	Medium: If not conducted properly, reported results may be off the 'true' results by multiple per cent.
Confidence	Low (the empirical evidence is based only on one workshop publication [43]).
Consensus	82% of the ACM RecSys Steering Committee agree with this best practice. *PLEASE NOTE: This is an example for illustration purposes. The percentage is made up.*

minor modifications. Initially, these guidelines do not require empirical evidence or consensus surveys. They can be simple and aligned with those used in the machine-learning community. Over time, these guidelines can be tailored more to fit recommender systems research, expanded and substantiated with empirical evidence and broader consensus.

The creation and justification of best practices can likely be undertaken by any motivated researcher with experience in recommender systems research. However, the final selection of these best practices, particularly concerning points 3 (severity) and 4 (consensus), should be conducted by reputable members of the RecSys community. This could be achieved through a Dagstuhl Seminar with selected experts or by the steering committee of the ACM Recommender Systems Conference.

In conclusion, establishing well-defined best-practice guidelines, endorsed by the community and enforced by key publication venues such as the ACM Recommender Systems conference and the ACM Transactions on Recommender Systems (TORS) journal, would be a significant move towards resolving the long-standing crisis in the recommender system

research community. For over a decade, the community has struggled with inconsistencies and lack of rigor in research practices. By adopting and enforcing these guidelines, we can ensure higher research standards, facilitate reproducibility, and contribute more robustly to collective knowledge.

4 Working Groups

4.1 Theory of Evaluation

Neil Hurley (University College Dublin, Ireland, neil.hurley@ucd.ie)
Vito Walter Anelli (Information Systems Lab, Polytechnic University of Bari, vitowalter.anelli@poliba.it),
Alejandro Bellogín (Universidad Autónoma de Madrid, Spain, alejandro.bellogin@uam.es)
Olivier Jeunen (ShareChat, United Kingdom, jeunen@sharechat.co)
Lien Michiels (University of Antwerp & Vrije Universiteit Brussel, Belgium, lien.michiels@uantwerpen.be)
Denis Parra (Pontificia Universidad Católica de Chile, dparras@uc.cl)
Rodrygo L.T. Santos (Universidade Federal de Minas Gerais, Brazil, rodrygo@dcc.ufmg.br)
Alexander Tuzhilin (New York University, USA, atuzhili@stern.nyu.edu)

4.1.1 Introduction and Scoping of the Problem

It is commonly believed that the "best practices" of empirical evaluations, as presented in the recommender system literature, are often unclear. The methodology used to carry out the evaluation is not clearly defined or is incomplete, and is not properly justified or aligned with the theoretical foundations of performance evaluation methodologies previously developed in the fields of machine learning and statistics. Therefore, the choice of metrics to compare performance across systems and the configuration of these metrics can seem arbitrary. In [39], for instance, it is argued that the way recommender systems researchers do evaluations, model selection, data splits and so on, is generally very poor with little consistency and no easy way to compare results. It should be a major wakeup call to the RS community to sort this out. This section aims to clearly articulate the deficiencies in current evaluation practices and to present a way forward so that evaluation can be improved in the future.

Evaluation of recommender systems has a very broad scope. There are many different types of recommender systems, from conventional top-N recommenders, to conversational recommenders, federated systems and reinforcement learning systems, to name just a few. Moreover, there is a great variety of aspects of recommendations the performance of which we may be interested in measuring. Some examples of these aspects include [48]:

- Ability to predict item relevance.
- Ability to rank items according to relevance.
- Novelty of recommended items.
- Diversity of the recommended set of items.
- Item coverage.
- Serendipity and unexpectedness of the recommendation.
- Fairness across users and items.

- Business oriented performance in terms of items clicked, adoption and conversion, the churn rates, sales and revenue, and other business performance metrics capturing consumer preferences and levels of consumption.
- Efficiency/latency of the recommender algorithm.
- Privacy of the system data.
- How explainable the recommendations are.

Evaluation in recommendation systems [48, 64, 71, 83] has been inspired by evaluation in machine learning [65, 28], information retrieval [5, 77, 79], and statistics [30, 68]. However, the assumptions behind those original procedures and metrics might not always hold in the context of recommender systems. For example, use of the nDCG or AUC metrics without proper justification can lead to biased or improper evaluations in some RS applications. There is a need to revisit the assumptions behind the original metrics and their suitability to the evaluation of recommendation systems. This section focuses mostly on metrics that are used to measure relevance or ranking performance in offline evaluation methodologies. It also examines the relationship of these metrics to online performance characteristics that they are used to predict.

The following deficiencies of evaluation methodologies as currently practiced are identified:

1. Evaluation protocols are usually chosen arbitrarily, without proper justification of their use and/or proper grounding in the previously developed evaluation methodologies in the fields of statistics and machine learning. A typical justification is often based on citing previous work that used the same protocol, which sometimes has its problems recursively leading to the "original sin" paper having various methodological issues.
2. The theoretical assumptions required to justify the choice of an offline metric are generally not known. The community needs to be made aware that certain metric choices carry an associated implicit set of assumptions about the problem context.
3. How statistical significance testing is carried out is often not clearly articulated, and it is generally not well enough known that particular statistical tests are based on assumptions about the data that may or may not hold. Researchers and practitioners need to be more mindful of the appropriateness of any test that is chosen.
4. Papers generally fail to report more than a summary performance statistic, averaged over the user population, rather than examining the dispersion of performance across the population, or the full distribution of the performance metric.
5. Related to the above, the uncertainty in the performance measurement is generally not reported.
6. It is difficult to introduce new performance metrics to the community and have them accepted and adopted.
7. Best practice in performance evaluation has been studied in a number of related domains, and, where appropriate, such best practice should be transferred into recommender performance evaluation. Related domains include:
   - Information retrieval,
   - Marketing – in which there exists extensive knowledge on how to calculate various marketing performance-based metrics and also properly carry out randomized control tests (RCTs). Some of this knowledge can be applied to recommendation problems [87, 56, 58].
   - Economics – in which econometrics-based models deal with controlled experiments that help to establish causal relationships in economics-related problems,
   - Applied statistics – in which some of the statistical methods have been applied to recommendation problems [2].

8. How many metrics and which metrics should be reported for a particular system is unclear. Some studies have shown strong correlations between some of the commonly-used metrics in certain contexts [38], while, on the other hand, when the size of datasets grows significantly, metrics can behave quite differently. This observation has been made across various fields of data science, including the insights made by Peter Norvig from Google back in 2010[2], and the field of recommender systems should examine this issue in greater depth.

In the rest of this section, we will discuss some of these issues in greater detail.

4.1.2 Data Preparation

This section delves into the crucial role of data preparation and its relationship with evaluating recommender systems. Three key preprocessing phases are explored: labeling, filtering, and partitioning. Each technique significantly impacts the evaluation process and requires careful consideration.

As a first step, often labeling takes place, where the quality of labels assigned to interactions (relevant vs not relevant) may directly affect how well a recommender system is evaluated. Choosing these interactions is critical, as it filters out others. This section discusses the implications of labeling and the challenges it presents.

As a second step, filtering is a preprocessing phase employed to achieve various objectives, including sparsity handling, noise reduction, accuracy improvement, and the alignment of the content information with interactions (also referred to as side information alignment).

Finally, a partitioning of the dataset is made to train the model. In fact, datasets used for recommender system evaluation are essentially samples from a larger, unobserved population, and to guarantee acceptable generalization capabilities for the trained recommender systems, data partitioning plays a crucial role. This section discusses various data partitioning strategies, along with their underlying assumptions and potential limitations. The section explores how these assumptions can affect the ecological validity of the evaluation, meaning how well the results translate to real-world scenarios [45].

4.1.2.1 Labeling

The quality of labeling significantly impacts a recommender system's evaluation. In this context, labeling refers to identifying which interactions are relevant to the system's goals. The choices made in this respect are pivotal and can skew the overall evaluation. Since choosing which interactions are relevant inherently filters out others, labeling acts as a filtering step (the following section discusses the implications of data filtering).

Compared to information retrieval, recommender systems deal with a much smaller portion of items actually examined by each user. As a result, the system designer lacks complete knowledge of which items are relevant to individual users. Further complicating matters, "*the non-observed user-item pairs – e.g. a user has not bought an item yet – are a mixture of real negative feedback (the user is not interested in buying the item) and missing values (the user might want to buy the item in the future)*" [67]. In recommender systems, data is often missing not at random, as highlighted in previous research [18, 61, 80, 81, 52]. Unlike other fields, where labeling the entire dataset is a condition to train the model,

[2] https://www.nyu.edu/about/news-publications/news/2010/september/google-research-director-peter-norvig-on-the-unreasonable-effectiveness-of-data-sept-17-at-courant-institute.html

this section focuses on strategies to handle limited feedback. The researchers should be particularly aware of this consideration. Indeed, previous literature showed that recommender system performance measured on a fully observable dataset substantially differs from the one computed on a partially observed dataset [17].

Given the limited feedback and the difficulty of collecting a reasonable amount of feedback, many approaches focus on unary feedback that is easier to collect [91]. Consequently, to assess these systems' performance in the case of multi-valued feedback (e.g., on a $1 \dots 5$ scale), the conversion to unary feedback is necessary. However, passing from multi-valued to unary feedback has important implications on evaluation since several metrics consider the "relevance grade" in the computation of the formula (e.g., nDCG). The interested reader may find a more detailed discussion on this topic in the following sections. How the conversion is performed is pivotal for the entire evaluation process. Moreover, each technique comes with its assumptions, whose absence hinders their applicability. This operation is usually performed by using

- **a global threshold**. A single threshold is defined at a global level (e.g., 3 on a $[1 \dots 5]$ range), and every rating above (or equal to) the threshold is considered a relevant interaction. If the practitioners adopt this approach, they are implicitly assuming that all users should have the same rating distribution, or at least, on a more psychological level, every user values in the same manner the various grades of the reference scale (e.g., for all users the value 3 should indicate a barely acceptable item). While this assumption is generally unwarranted, the extent to which it might hold could depend on whether an explicit meaning has been assigned to each grade in the user interface at rating time.
- **a per-user threshold** (user rating average or median). A single threshold is defined based on user-specific characteristics, like their rating distribution. A common approach is to consider the user ratings' average as the threshold. However, this approach brings an even stricter assumption: each user must have a balanced distribution of ratings between positive and negative feedback. Otherwise, the semantics of the threshold cannot match (e.g., if a user only rated positive items, using the mean or the median of the ratings to define the threshold is meaningless).

4.1.2.2 Filtering

The filtering step is an important preprocessing phase aiming to achieve one or more objectives, including:

- Sparsity handling: It can be particularly helpful for recommender systems dealing with sparse data, where some users may have not interacted with a sufficiently large portion of the items.
- Reduce noise: By excluding sparsely connected users and items, k-core filtering can minimize the impact of noisy or irrelevant data points on recommendations.
- Improve accuracy: Focusing on denser user-item relationships potentially leads to more accurate recommendations because the system is considering stronger user preferences and item connections.
- Side-information alignment: It could be necessary in case of comparison of a collaborative filtering method with a content-based or a hybrid model. It ensures a fair comparison between the different families of recommendation algorithms.

There are several techniques devoted to filtering datasets. In this section, some of the most adopted ones are briefly discussed:

- **User interaction threshold filtering.** This approach filters users based on their overall interaction volume with the system. The system designer sets a minimum (or maximum) number of interactions a user must have to be considered for recommendations.
- **Item interaction threshold filtering.** This approach filters items based on their overall popularity. The system designer sets a minimum (or maximum) number of transactions an item must have to be considered for recommendations.
- **K-core filtering.** It combines the two previous approaches and identifies a denser user-item network. It first builds a network where users and items are nodes while interactions are edges. Then, it **iteratively** removes users or items with less than a chosen threshold (k) of connections. This creates a k-core, a subnetwork containing only well-connected users and items. Recommendations are made based solely on this core. Unfortunately, if the filtering procedure is not repeated until convergence (until no more users or items are removed from the network), the k-core subnetwork is not created, and some users and items may have an uncontrolled number of interactions, making the overall evaluation unfair and not replicable.
- **Content (a.k.a. side) information alignment.** When the experimental evaluation comprises models that leverage content information (e.g., images, categorical or numerical features, graphs, semantic information, textual descriptions), the alignment of interaction information with content information is necessary. Suppose a researcher proposes a visual-based recommendation method that exploits images of the products. However, only 50% of the items contain visual information. The choice of not aligning the interactions with the content information will result in an unfair comparison, and the quality of the proposed recommender could not be assessed. Lastly, side information alignment impacts both users and items since, after the filtering, some users could have an empty interaction history and will be removed.

The main problem, hence, is that *after applying any filtering approach, the distributional characteristics of the dataset are different.* Depending on the degree of change in dataset characteristics, the new dataset may no longer reflect the original dataset, thereby undermining the internal and ecological validity of the experiment. In case the practitioner is going to use the learned model in a real world production environment, there are no theoretical guarantees that the model is going to perform as intended.

Nevertheless, if the learned recommender system is not going to be used with the original (unfiltered) data, filtering could be justified and used since it creates a (different) new dataset potentially useful for research purposes. It is worth mentioning that the extent to which this new dataset is realistic is outside the scope of this document. Whatever the rationale behind the researchers' choice, every time they apply any filtering method, they should report the new dataset characteristics. Some widely employed dataset characteristics considering distributional and topological properties are [1, 3]:

- Rating space structure:
  - **Size of rating space.** The size of the rating space can be computed as $RatingSpace = |U| \times |I|$;
  - **Shape of rating space.** The shape of the rating data is captured by the ratio of the number of users and the number of items in the rating data. The shape of a rating space is captured by the user-item ratio, that is, $UserItemRatio = |U|/|I|$.
  - **Rating density.** This can be calculated as the proportion of known ratings (i.e., provided by the users to the system) among all possible ratings that can possibly be

given by the users. More specifically, $density = |R|/(|U| \times |I|)$.
where U, I, and R indicate the sets of users, items, and interactions, respectively.
- Rating frequency distribution:
  - **Basic shape.** The basic shape of the frequency distribution of user or item ratings can be computed by using the first four moments: *mean, variance, skewness, kurtosis.*
  - **Concentration.** The concentration of items or users in the frequency distribution can be calculated by using inequality measures including Gini coefficient, Pareto exponent, Simpson diversity (or Herfindahl) index, and Shannon's diversity index (or entropy).
  - **Average user degree.** It refers to the average number of interactions per user.
  - **Average item degree.** It is computed as the average number of interactions per item.

Finally, it is essential to underline that the dataset characteristics should be reported to characterize the new dataset and not to claim that the system trained on the new system will perform comparably on the original dataset. Indeed, preserving the dataset's statistical properties is insufficient to guarantee similar performance. The consideration paves the way for the open challenges in data modeling and simulation for recommender systems. The reader may find a detailed discussion on the relationship between simulation and evaluation in Section 4.1.6.

4.1.2.3 Partitioning

The datasets we use to evaluate recommendation algorithms are essentially an observed sample, sampled from an underlying, unobserved, true distribution of data. To avoid the model overfitting a specific observed dataset when we evaluate the performance of a recommendation algorithm, the dataset is split into separate training and test datasets, where the former is used to estimate the model and the latter to evaluate its performance. If there are hyperparameter values to be estimated as well, the best practice is to create yet another separate dataset, the validation dataset, that is used solely for the purpose of determining the optimal[3] hyperparameter values.

There are several "best practices" in use for partitioning datasets. Most partition data per user, i.e., some portion of a user's data is assigned to each of the training, validation and test dataset. For example, the predominant strategy for data partitioning splits the entire data in a "random" fashion, by assigning some of a user's interactions to each portion – training, validation, and test – uniformly at random [83]. In practice, the training dataset is typically selected to be many times larger than the validation and test datasets. Such a partitioning of the dataset is stochastic in nature: Different random seeds will result in different partitions of the dataset. As such, uncertainty can be decreased by repeatedly splitting the dataset in this fashion for different initial values of the random seed, a procedure typically referred to as "(Monte Carlo) cross validation."[4] While a standard in many other fields of machine learning [37], cross validation is rarely performed in the practice of offline evaluation because it significantly increases the runtime and cost of the experiment. As a result, offline empirical studies frequently report on very uncertain and biased point estimates of a recommendation algorithm's performance. In addition, while the theoretical basis of cross validation has been studied for other domains of machine learning, e.g., classification [37], it has not been studied in the context of recommender systems, to the best of our knowledge.

[3] Optimal is taken to mean "optimal for the validation dataset" here, which need not have a relationship to true optimality.
[4] This is just one example of a cross validation strategy.

Partitioning data by selecting samples, i.e., user interactions, to be assigned to each of the datasets uniformly at random ensures that the training, validation, and test datasets have similar distributional characteristics. As a result, the recommendation model, estimated based on the training dataset, will likely be a reasonable model for the test dataset and the observed dataset as a whole. However, it also makes several strong assumptions about the real world phenomenon that we are trying to estimate by means of the recommendation model, which may undermine the model's applicability in the real world and the experiment's ecological validity.

Firstly, it makes the assumption that there is no inherent ordering to a user's interactions. While this may be a valid assumption in very specific cases, it cannot be assumed to hold in general. We can make several simple counterexamples for popular practical use cases of recommender systems. For example, in e-commerce, users are highly unlikely to purchase a game for the PS5 or Nintendo Switch, if they do not own the appropriate gaming console.[5] In movie recommendation, users are unlikely to watch the first installment of a series, after watching the second and third.[6]

Secondly, it makes the assumption that a user's interactions are independent of time, or, in other words, static. It is easy to see how this assumption too may not hold in the real world: in real world recommender systems, new items are introduced frequently, whether they be new books, new music, new movies, new articles, new applicants and new jobs, ...

Both phenomena have received some attention in the literature, and alternative data splitting strategies have been proposed that do not make one or both of these assumptions. An "order-aware" or "user timeline" data split lifts the assumption that a user's interactions are unordered, and splits them so that the user's earlier interactions are used to predict their later interactions [53, 47]. A "time-aware" or "global timeline" data split lifts both assumptions by partitioning the dataset based on a timestamp, such that all interactions before this timestamp are assigned to the training dataset and all interactions after are assigned to the test dataset. While these data splitting strategies may alleviate the issue of ecological validity of an offline experimental result, they introduce yet other issues. Firstly, depending on the degree of data drift in the dataset, the training and test datasets may come from different data distributions, and as a result, the trained model may no longer be a reasonable estimator. Secondly, both methods are deterministic, i.e., provided that the timestamp or amount/ratio of interactions to assign to the test dataset is known, they result in a single unique split and thus a single, biased and uncertain, estimate of the trained model's performance. Practical strategies have been proposed to address this, e.g., cross-validation-through-time [54], sliding-window-evaluations [47, 52], or the timeline scheme [53]. However, these strategies lack theoretical foundations: it is unclear whether or not they result in less biased estimators of performance.

Authors of empirical studies that employ offline evaluations do not typically justify why the above assumptions may be assumed to hold for the specific observed dataset(s) that they are using. To the best of our knowledge, there is no theoretical basis for these assumptions, nor knowledge of how violating these assumptions may affect experimental results and the ecological validity of our offline evaluation experiments.

[5] Unless the game is purchased as a gift.
[6] Unless the series in question is Star Wars, or Marvel.

4.1.3 Configuration of Metrics

As surveyed in [4], ranking metrics are the most popular type of metrics being reported nowadays in the recommender systems literature. In this context, which cutoff – i.e., up to which N position the top-N elements in a ranking list are considered – is selected is an important decision. While it is acknowledged that this decision should be tailored to the task at hand or related to the interface the actual system was being or will be tested in [60], however, no such justification is found in many publications. This decision is not trivial, since, as analyzed by [88], some cutoffs might provide more robustness in terms of incompleteness (to sparsity and popularity biases) than others; in particular, longer cutoffs are more robust, even though the correlation between the obtained results was in general very high (above 0.90).

At the same time, results from metrics are reported in combination with other metrics [4]. As observed by [38], there is strong correlation (measured as "linear agreement" in the original work) among different subsets of popular evaluation metrics. More recently, a wider range of metrics was analyzed by [88] and consistent results were obtained. Hence, there is little gain in reporting metrics that measure very similar signals; instead, complementary measurements should be preferred, such as providing diversity or novelty metrics together with accuracy.

An even more critical aspect to be configured when dealing with ranking metrics is the process known as candidate item selection or sampling [22]. Here, the designer needs to decide which items should be requested to rank for each user. As discussed by [38], if we assume that the distribution of relevant items and non-relevant items within the user's test set is the same as the true distribution for the user across all items, then computing our metrics on the items for which we have ratings – i.e., the user's test – will be much closer approximations of their true values. However, this is the opposite of what is typically done in information retrieval, and in fact it does not mimic a realistic scenario where the user's test is unknown.

The impact of this specific decision was analysed by [10], evidencing that the former design obtains results consistent with error metrics, whereas the latter is the most appropriate one for ranking metrics. This effect is linked to how much unknown relevance is added to the test set, leading to the so-called *sampled metrics*, where a parameter is considered for the amount of sampled unknown (and, hence, non-relevant) items are included as candidate items to be ranked by the algorithms. Despite the potential benefit of using this configuration in the metrics because of the reduced computational cost (since not all items need to be ranked anymore), it has been found in several works that doing this may lead to inconsistent results, depending on the parameter considered and the dataset [55, 57, 18]. Nonetheless, it is interesting to consider that this sampling could be exploited to alleviate popularity or sparsity biases, as done in [42, 11]. Hence, this might be a potential avenue to be explored in the future, so that the impact of this configuration is analysed from other perspectives, such as multi-stakeholder or long-term evaluation.

Finally, there are other decisions more related to the technical details or implementation nature of the evaluation metrics that deserve a formal justification or, at least, as much transparency as possible from the researcher perspective, to properly assess the validity of the results obtained through the presented evaluation [10]. On the one hand, some metrics present different variations in the literature, each entailing a different assumption with respect to the user behaviour or the meaning of the results. A paradigmatic example is the configuration of nDCG [46], which requires a discounting function and a weighting scheme to transform the ground truth into relevance weights. While in the original paper the authors

discuss the underlying consequences of using 2-3 variations for these parameters, to the best of our knowledge there is no thorough study to understand the impact of each of these variations (or any other whatsoever) on the recommendation problem. In fact, it might be possible that, depending on the user task, domain, or additional constraints, one variation might be more adequate than another.

Similarly, how the relevance score is obtained from the information included in the test set, is sometimes not explicitly mentioned, and it is even difficult to determine from public implementations [10]. This is especially important for the cases where a rating scale is available in the dataset, whereas the evaluation metric expects either binary relevance (relevant vs non-relevant) or graded relevance (how each rating maps to the different relevance levels). Related to this issue, how the evaluation metrics are configured when a recommender provides a shorter list than expected – i.e., shorter than the provided cutoff – may make a great difference on the reported results, and more importantly, on the hypotheses being assumed as a consequence of that decision [10, 22], even coining the term *coverage shortfall* [19]. Let us take the example of the recall metric, which takes the number of relevant documents recommended up to position N and divides, in its original formulation, by the number of relevant documents known in the ground truth of that user [38]. It is straightforward to observe that, in some cases, it will be difficult to achieve a value of 1 at high positions, since N might probably be smaller than the size of the user test. To address this, some works such as [59] proposed a formulation that normalises by the minimum between the size of the user test and N; in that case, it might be possible to achieve the maximum value of the metric, even when the recommender has not ranked all the items the user has in their test set. By doing this, two orthogonal evaluation dimensions are being assimilated: recall and coverage; it is now impossible to discern a recommender that provides N *good* recommendations (in a ranking of size N) from another that only provides 1 recommendation matching the user test.

Moreover, "*matching the user test*" is also configurable and justifications about whatever decision made should be explicit and aligned with the problem at hand. Usually, matching the user test corresponds to recommending the exact same item the user has in their test set. However, whenever the domain is too sparse or there are obvious similarities between the items, researchers have considered some kind of similarity within the evaluation metric to discriminate between recommendation algorithms, by claiming that not all the recommended items are equal to each other, but some are better (and actually perceived better by the users) than others [31, 76]. This shift from exact to similar matching must be made crystal clear when reporting the results, as it may artificially boost the performance values, even at the expense of losing discrimination power, for example by using a similarity metric that is too vague.

4.1.4 Theoretical Justification of Offline Evaluation

Recommender systems are inherently targeted towards real-world end users, and their goal is often framed as trying to maximise the utility that these end users can get from the recommendations. This "real-world performance" is the estimand we care about in any evaluation procedure, be it online, offline, simulated, or measured via user studies.

Online evaluation is costly and requires access to end users, simulations require assumptions that are often hard to motivate or validate, and user studies take time and are well-suited for a limited set of research questions. Partly because of these reasons, offline evaluation is the most common paradigm in the research literature on recommender systems – and also commonly used by practitioners to obtain initial performance estimates. Broadly speaking, the goal of any offline evaluation procedure is then to estimate this "real-world performance" as best we can, in a reliable, reproducible, and robust manner.

Problematically, the community has repeatedly reported mismatches between offline results and real-world utility for more than a decade [66, 8, 52, 47, 7, 33, 69]. It is our belief that the theoretical disparity between commonly used offline evaluation procedures and metrics is at the heart of this: recall and (n)DCG are well-motivated in general machine learning (ML) or information retrieval (IR) settings respectively, but the assumptions required to justify their use are rarely mentioned explicitly in recommendation research. Assumptions permeate our scientific field, and some are easier justified than others. Being explicit about them provides clarity about the limitations certain methods have, and hints at potential ways forward: "*Can we lift these assumptions? Can we quantify the bias on the estimator that is a result of violated assumptions? Which set of assumptions is necessary and sufficient for a metric to be theoretically justified?*" With the prevalence of offline evaluation, finding answers to these questions is crucial. Nevertheless, we find that such questions are rarely posed in the first place, and the motivation for specific evaluation metrics boils down to matching the recommender systems problem to either ML or IR. Whilst clearly related, there is *no* exact match between typical applications in these settings, and any procedures and metrics we bring into the field should be vetted as a result.

[46] introduce the (normalised) DCG measure in the context of classical information retrieval applications, like web search. They write: "*a simple way of discounting [...] is to divide the document score by the log of its rank*". It is clear that this proposed discount function was effective, and it has been adopted by the IR and consequently by the RecSys communities. Nevertheless, the choice of discount function carries implicit assumptions about user behaviour, and how they interact with a ranked list of recommendations in terms of examining items. "*Simple ways*" can be intuitive, but deeper theoretical justifications allow us to formally link offline evaluation measures to online metrics we might care about. [15] proposed similar metrics that leverage an estimate of the probability that a user will see a recommendation in a ranking – and it should be clear that the accuracy of that estimate affects the utility of the evaluation metric. Indeed, recent work reports that improved exposure probability estimates improve correlation with results obtained through online experiments [49].

One approach to connecting an offline measures with real-world performance is to demonstrate that the offline measure is an unbiased estimator of the online performance characteristic that is ultimately of interest. This is typically framed as "counterfactual" or "off-policy" evaluation, and has gained traction in recommendation applications [90, 72]. Several studies have reported that careful application of such techniques can close the gap between offline evaluation results and real-world performance as measured in an A/B-test [34, 35]. Nevertheless, it is often seen as a "niche" area of research, and connections to evaluation metrics that are prevalent in the field are unclear.

In an attempt to close this gap, [51] examine the assumptions of the problem context that are required for Discounted Cumulative Gain (DCG) to be an unbiased estimator of "online reward". Broadly, these include that the reward for an item is independent of past recommendations (avoiding the need for reinforcement-learning-type evaluation); that the probability that a user views an item at a particular rank depends only on that rank, and not on any actions taken on items in other ranks; and that the reward is independent across all ranks. One way, therefore, to justify the choice of DCG for offline evaluation is to argue that the problem context satisfies these assumptions. Nevertheless, such assumptions are not generally known to practitioners or researchers – even for the methods and metrics that underpin our research field that is largely driven by empirical progress. *Further theoretical analyses that identify connections between offline metrics and real-world performance are to be encouraged in the community.*

It is also worth noting that other common metric constructions, such as normalising by the DCG of an "ideal" ranker, when no such ideal can be determined, or normalising before averaging, undermine the theoretical link between the estimator and the estimand. As a result, they can change the order in which system performances are ranked without any sound justification for modifying the measure in a way that moves the optimum. Anecdotal evidence seems to imply that this consequence is not widely known, which is troubling given the prevalence of the procedure.

Normalised DCG is a staple for evaluation in the IR community, and this has motivated its use in recommender systems research. Typical IR applications like web search rely on datasets that have some "ground truth": often these are relevance judgments collected from experts. In recommender systems research, the very nature of the problem setting inhibits us from acquiring anything like this.

Part of the problems mentioned above regarding nDCG stem from its unrealistic setting with respect to partial information, which is prevalent in our community. In fact, the original article [46] claims "*they (nCG and nDCG measures) represent performance as relative to the ideal based on a known (possibly large) recall base of graded relevance judgments*". Hence, one of the underlying assumptions made explicit by the authors is that *the normalisation should be done on a large recall base of ground truth* or, in other terms, that unless ground truth is large enough, we would not have enough confidence on the "ideal" value of the metric. This extends to other recall-oriented metrics, like Recall (obviously in its original formulation [36, 38] or in recent normalisation variations [59]) or Mean Average Precision (MAP) [5, 38]. Here, the main assumption being violated is that, usually (unless the full user-item interaction matrix is known), in recommender systems ground truth is far from complete, hence these metrics are being computed under a wrong premise: that the observed preferences is what the recommender system should achieve, ignoring that these are a minor representation of the real user preferences.

Even though this problem is not as severe in the information retrieval area, there are proposals aiming to tackle this issue. For example, the *bpref* metric [16] was specifically defined to be robust to incomplete judgments sets; however, it is seldom used in recommendation tasks [4]. Similarly, the variations of Average Precision presented in [95] (induced, subcollection, and inferred) provide robust measurements to incomplete and imperfect relevance judgments. Hence, the community should aim at understanding how to adapt these metrics to the recommendation domain, as in [89], and decide whether these are enough to address the aforementioned problems or if more specific measurements are needed.

To satisfy unbiasedness according to the derivation from [51], the discount function used in DCG should accurately reflect exposure probabilities. A common user model assumes that users decide whether to abandon the recommendation list with a fixed probability after every item. With this user model and the appropriate discount function, DCG becomes equivalent to the Rank-Biased-Precision (RBP) metric, which is commonly used in IR. [63] write: "*A useful consequence of the proposed RBP metric is that it is possible to compute upper and lower bounds on effectiveness, even when the ranking and relevance judgments are partial rather than comprehensive.*" Whilst less common in IR, as we have argued, incomplete relevance judgments are ubiquitous in recommendation use-cases. As a result, this insight is **crucial** for our community, as it hints towards ways we can quantify the statistical biases that arise due to violated assumptions. *Further theoretical analysis of such properties is an important and promising research direction.*

Specifically, bounds for more general discount functions that are, e.g., personalised and context-dependent, would be of both theoretical and practical importance. Indeed, if other covariates exist that impact exposure probabilities, we need to account for them to avoid problems of unobserved confounding that would inevitably lead to further biases [50].

So far, we have argued in favour of more rigorous theoretical justification of offline evaluation metrics and procedures, so we can make mathematically meaningful statements about estimates of real-world performance without requiring on end users that interact with the recommender system. Online evaluation procedures, on the other hand, leverage interaction with end users to directly measure the quantities we care about – be it short-term, long-term, multi-objective, multi-stakeholder, accuracy-, diversity-, or fairness-oriented. A/B-tests are typically used for this, because of their strong theoretical connections to well-known and well-vetted experimental setups like Randomised Controlled Trials (RCTs) [70, 50].

In line with the offline evaluation procedures we tend to borrow from ML and IR without questioning their assumptions, we analogously rely on the seminal works of [29] and [70] to motivate why RCTs and A/B-tests are the gold standard for measuring real-world performance. These methods were, nevertheless, originally proposed in different contexts, relying on different assumptions. This inhibits their direct application to recommendation problems, but the mismatch is rarely acknowledged in the research literature. [48] discusses problems that arise with machine learnt models that update over time: when training data is influenced by the treatment, the Stable Unit Treatment Value Assumption (SUTVA) is violated, undermining the credibility of the experimental setup. Similar observations have been made in industry settings, where bias and interference complicate reliable measurement of performance [85]. [6] focus on multi-sided experiments, where we, e.g., have item consumers and providers that can interfere and complicate statistical inference – a setup that describes most commercial instances of recommender systems. [82] propose specific adaptations to online evaluation procedures that minimise this type of interference, with a focus on "exploration". Notwithstanding this, interference also occurs even in simpler settings where we only consider users that can interact [20]. [50] provide guidance for online experimentation in general, describing common situations where problems can occur. These issues should be acknowledged and widely known, to avoid blindly putting A/B-test results on a pedestal as the "gold standard", without being clear about the assumptions.

[32] famously criticizes common mistakes in IR evaluation, some of which directly map to RecSys use-cases too, whilst other do not. Their criticism has been the subject of discussion itself, with [75] retorting some of the arguments and highlighting that there are differing theoretical views on evaluation in general. Such public discussions are healthy for the research community, and it is our belief that RecSys-focused extensions can be helpful.

4.1.5 Reporting Results

Most empirical research on recommender systems aims to introduce new methods or test existing methods in new applications by conducting experiments on one or more datasets. Properly reporting results is crucial for drawing robust and widely applicable conclusions about the proposed method, system, or application, especially in comparison to previous works. The current practice for reporting performance in the area follows a pattern: i) indicate a set of performance metrics (nDCG, MAP, recall@k, precision@k, AUC, diversity, novelty, etc.), ii) indicate a set of competing and baseline methods, iii) report the average of those metrics over multiple users, iv) in many cases, but not always, report some statistical test, v) in some cases, provide plots to visualize the behavior of the metrics as a function of hyperparameters or training variables. Although it seems like following this procedure ensures strong evidence to support conclusions, several assumptions behind this process can diminish their robustness.

Recent research on the evaluation of recommender systems is shedding light on reporting aspects that can strongly influence the final interpretation of the results. For instance, [51] studied the suitability of reporting nDCG to compare the performance across methods

since some of their assumptions are violated in recommender systems. [27] advocate for using distributions rather than only reporting point estimates of metrics to compare the performance of different methods. Moreover, research in recommender systems has stuck to reporting performance with a rather small set of metrics [4], whereas researchers in information retrieval have explored further to account for important aspects of ranking, suggesting the use of Expected Reciprocal Rank (ERR) [23], and Rank Biased Precision (RBP) [63].

Other fields have partially addressed these issues by continuously researching the assumptions behind evaluation metrics or directly introducing guidelines. For instance, the fields of Human-computer interaction and Information Visualization have advocated for reporting results using informative charts with effect sizes and interval estimates [26, 13, 84], rather than relying exclusively on p-values. The main criticism about the practice of only reporting p-values is the promotion of dichotomous thinking, i.e., the classification of statistical evidence as either sufficient or insufficient, typically through the use of arbitrary cutoffs such as the p-value $p < 0.05$ [12]. There is also wide consensus among statisticians about moving beyond p-values to advance research in general [68]. Moreover, the field of information retrieval has a tradition of continuously researching evaluation metrics and practices, with several tutorials and books over the years emphasizing guidelines and best practices [74, 62]. For instance, to assess for significance in information retrieval, [79] analyzed the robustness of several statistical tests and concluded that Wilcoxon and sign tests should be discontinued. This work has been continued with reports emphasizing a better understanding of statistical tests [21] and good practices to report significance beyond p-values [73]. Going further, they have expanded this research to online evaluation [40].

These are just examples of the need to revise the assumptions and procedures for offline evaluation that the recommendation systems community considers in the form of providing evidence of progress in the area.

4.1.5.1 Beyond Averages

Ensuring that evaluation metrics are aligned with the actual success criteria is a crucial first step towards assessing the effectiveness of a recommender system. In practice, the implementation of these metrics must address several data quality issues. As discussed in Section 4.1.2, for retrospective evaluation scenarios, this includes handling the available feedback as incomplete, noisy, and often biased samples of user behavior. In any case, for a robust evaluation, the effectiveness of a recommender system must be assessed across multiple users as test samples. On the other hand, summarizing per-user estimates through simple averaging fails to capture important aspects of the underlying effectiveness distribution across the entire sample, which is key to comparing systems.

Statistical Significance. Comparing averages may hide subtle yet important differences between systems. For instance, measured average improvements might come from only a handful of users in the test sample, when the majority of users might experience a decrease in their experience with the system. Such a variability in performance across users can be quantified to serve as an estimate of the uncertainty associated with the reported averages in the form of a confidence interval. Taking a step further, statistical hypothesis testing can be employed to quantify the extent to which the differences observed between systems are significant.

Despite being common practice in related fields [73, 86], significance testing is not as widely adopted in the recommender systems community [22]. Moreover, which testing procedure to use for different recommendation problems is often unclear. Recent results have shed

light on the statistical power of existing procedures when applied for typical recommender evaluation scenarios with sample sizes in the order of thousands of users [43]. In contrast to small-sample regimes typical to evaluation efforts in related fields (e.g., search evaluation campaigns with a couple hundred queries as test samples [92]), large-sample regimes render existing significance testing procedures robust to violations of their underlying assumptions (such as normality and homoscedasticity [86]). In this scenario, having a significance testing procedure in place becomes more important than which particular test to choose. Another relevant aspect to consider when assessing statistical significance is the increased probability of falsely detecting significant differences (aka Type I errors) stemming from the simultaneous comparison of multiple systems [44].

Practical Significance. Statistical significance tests can help detect unpromising recommendation approaches early on in the process of searching the space of effective solutions. Nonetheless, a statistically significant improvement may not necessarily be of practical significance for the recommendation scenario under consideration. In particular, confidence intervals are a function of both the effect size – the magnitude of the improvements observed with respect to a baseline system – and the sample size – the number of observations – when comparing systems. Therefore, reporting effect sizes is of utmost importance for assessing the practical significance of a result. Indeed, regardless of its magnitude, a positive effect size indicates a consistent improvement across users in the test sample. Depending on the target scenario, even a small – yet positive – effect size may be of practical significance, considering the scale involved (e.g., a tiny increase in revenue per user across a large fraction of the user population).

While positive effect sizes indicate a consistent improvement, they do not tell the full story. For instance, data incompleteness issues often lead to a very low (if not zero) performance for many individual users, which may severely affect the measured average performance of different systems or even the effect size when comparing systems. Inspecting the underlying distribution of improvements across test users may reveal important insights into the relative strengths and limitations of the systems being compared. Indeed, looking at performance differences at an individual level could help mitigate the risk of deploying a new system that brings average improvements at the expense of hurting the experience of several individual users. Moreover, segmenting test users according to some discriminative user feature (e.g., demographics, past interests) may help surface inherent difficulties of the systems or even an unfair treatment against certain user groups [27].

Other Considerations. In addition to assessing the statistical and practical significance of the reported results, other effects are also worth analysing when evaluating recommender systems. One such effect is the sensitivity of a recommender system to its hyperparameters. Given the costs involved in hyperparameter tuning, particularly for compute-intensive systems deployed in massive-scale recommendation scenarios, understanding the extent to which the effectiveness of a system depends on the configuration of each of its hyperparameters may lead to more cost-effective deployments. Moreover, understanding the impact of different components of the system on its final performance through an ablation analysis can be also informative. Indeed, not only does it help determine the cost-effectiveness of each component individually, but also to narrow down the cause of the observed improvements. The latter can be of particular importance as means to identify the actual scientific progress brought by each newly introduced approach.

Lastly, as in every scientific undertaking, clearly reporting the limitations of the conducted experiments is crucial for several reasons. Transparent reporting allows other researchers to accurately interpret the results and understand the context in which the findings are

applicable. It helps identify the potential sources of bias or error that may have influenced the outcomes, such as sample size limitations, data quality issues, or specific assumptions made during the analysis. Acknowledging these limitations also facilitates reproducibility and comparability, enabling other researchers to replicate the study under similar or varied conditions to verify the findings. Furthermore, it guides future research by highlighting areas that require further investigation or improvement, thereby contributing to the overall advancement of the field. Clear communication of limitations fosters trust and credibility in the research community and ensures that the conclusions drawn are robust and reliable.

4.1.6 Data Modeling, Synthetic Data Generation and Simulation

It is common practice to evaluate an algorithm over a number of empirical datasets in order to demonstrate its performance. The measured performance is intimately connected to the characteristics of these datasets, so that in order to truly test the applicability of the algorithm, it should be evaluated on as wide a range of datasets as possible, covering a range of characteristics. There are a number of recognized challenges to this approach, such as ensuring that the chosen datasets cover a sufficient range of interest. Common preprocessing steps can change the distribution of the data, as discussed elsewhere in this section. Moreover, evaluation over a data snapshot from a live system assumes that this dataset contains sufficient information to determine future user behaviour and this may not indeed be true.

A more statistical approach is to define a data generating distribution and to evaluate performance as a function of the parameters of that distribution. In doing so, one could in theory explicitly explore the relationship between the data characteristics, as controlled by the parameters, and performance. Parameters can be adjusted beyond the range that might be available in empirical datasets, so that, for instance, exploring an algorithm as the number of users or catalogue size is scaled upwards, or over different levels of sparsity, becomes possible. Additionally, evaluating over a data distribution or over synthetic data drawn from a data generator, has the advantage that it avoids issues of privacy and limited access to real-world datasets.

Going further in this direction, a full user behavioural model can be proposed and implemented in a simulator. The recommender system algorithm is then evaluated against the parameters of the simulator, which can be modified to explore different user behaviours. It is noteworthy that as far back as [38] simulation is mentioned as a means of generating training data. However, these purely theoretical approaches (in so far as they do not rely on real-world datasets), are not commonly adopted in the community because of the inherent challenges of selecting statistical distributions that cover the characteristics observed in real-world data and of modelling real user behaviour. As a result, they have not gained much traction as valid evaluation methodologies to date.

A full review of data generation techniques and simulation is beyond the scope of this report. We mention some work as follows:

- The impact of data characteristics on recommender system performance has been explored in [1, 25, 24], where the identified characteristics include rating distribution, as summarised by a few moments of the distribution, data sparsity, user and item rating frequency distribution.
- A number of models for synthetic dataset generation have been proposed, notably models based on fractal expansion [9], models based on generative models such as Variational Autoencoders (VAEs) [93] or Generative Adversarial Networks (GANs) [14, 78].

- Work on simulators for recommender systems is mostly in the context of reinforcement learning approaches, in which it is difficult to obtain real-world data to evaluate long-term reward. Simulators include RecSim [41], RL4R [94], KuaiSim [96].

In the current state-of-the-art, the following gaps may be identified. The measures used to characterise datasets are not sufficient to determine algorithm behaviour. Capturing the correlations between interactions is difficult to model and cannot be easily captured in a small set of summary statistics. Synthetic data generation methods can prove very useful, in particular in avoiding issues of privacy and regulation associated with the use of real-user data, but do not allow for full control over the dataset characteristics that we may want to explore. Much work is required in user modeling in order to develop simulators that can accurately model real-user behaviour, across a range of different recommender system contexts.

We argue that further work in these directions can be very valuable to developing more robust evaluation methodologies that are not dependent on the availability of empirical data.

4.1.7 Practical Issues of Improving Evaluation Methodologies

While we strongly advocate for improvement of evaluation methodologies over the current common practice, it is important to recognize the practical issues raised by committing to a robust evaluation protocol. We can certainly learn from the medical community, for example, in terms of adopting standardized reporting styles that concisely capture the details of the statistical analyses that have been applied. Nevertheless, applying the rigor of a very strong analysis protocol inevitably means that the time dedicated to evaluation becomes significantly longer. Moreover, the computational resource required to fully evaluate an algorithm over a range of settings is substantial, considering the training times associated with deep models that are more and more often being adopted. The environmental impact of such analyses needs also to be considered.

The primary focus of our community is the development of novel models and algorithms. The resource commitment to evaluation will detract from this primary focus and slow technological advancements. The scope of recommender systems is very broad nowadays, beyond their original application in e-commerce, to systems applied in health domains. One view is that we need to weigh the cost of rigorous evaluation against the cost of an erroneous assessment of an algorithm's performance. At one extreme, a poor assessment may do no more than slightly dis-improve user experience with a non-critical application, while on the other it may have significant financial impact on a company that deploys an algorithm under false expectations or even be critical to the health and well-being of people to whom recommendations are made. Where substantial financial cost or cost in human life is at stake, then it is essential to do full and thorough assessment. For less critical applications, we may be content to observe an algorithm's true performance, once it is deployed in the wild.

As things stand, researchers spend so much time worrying about the validity of performance results in the state-of-the-art, so that systems have to be repeatedly re-evaluated for each new experimental comparison. We need to at least reach a point where experiments are clearly and fully reported, including the assumptions that go into the evaluation methodology, so that future researchers can rely on the soundness of the results and not feel obliged to repeat the analysis.

4.1.8 Recommendations

We summarize some broad takeaways from the above discussion. It is evident that some of the issues identified in this section are generic issues for data analysis and machine learning in general. Others are specific to the recommender system context in particular, and we need to be particularly mindful of circumstances in which findings established in a different domain are adapted to the recommender system domain.

- Be aware of the theory underlying evaluation that is already known and put it into practice.
- Be aware that assumptions underlie all evaluation choices and be conscious of those assumptions.
- The community should carry out further exploration of the theory developed in other disciplines and its adaption to the recommender system context, taking into account the specific characteristics of our domain.
- Further research on the theoretical grounding of data partitioning, labeling and filtering is necessary.
- Further theoretical analysis that identifies connections between offline metrics and real-world performance is required.
- Further research on improved models of datasets, user-modeling and simulation can alleviate the reliance of evaluation methodologies on the availability of empirical datasets.

References

1 Gediminas Adomavicius and Jingjing Zhang. Impact of data characteristics on recommender systems performance. *ACM Trans. Manag. Inf. Syst.*, 3(1):3:1–3:17, 2012.

2 Deepak K. Agarwal and Bee-Chung Chen. *Statistical Methods for Recommender Systems*. Cambridge University Press, 2016.

3 Vito Walter Anelli, Daniele Malitesta, Claudio Pomo, Alejandro Bellogín, Eugenio Di Sciascio, and Tommaso Di Noia. Challenging the myth of graph collaborative filtering: a reasoned and reproducibility-driven analysis. In Jie Zhang, Li Chen, Shlomo Berkovsky, Min Zhang, Tommaso Di Noia, Justin Basilico, Luiz Pizzato, and Yang Song, editors, *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023*, pages 350–361. ACM, 2023.

4 Christine Bauer, Eva Zangerle, and Alan Said. Exploring the landscape of recommender systems evaluation: Practices and perspectives. *ACM Transactions on Recommender Systems*, 2(1), mar 2024. URL `https://doi.org/10.1145/3629170`.

5 Ricardo Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval – the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England, 2011.

6 Patrick Bajari, Brian Burdick, Guido W. Imbens, Lorenzo Masoero, James McQueen, Thomas S. Richardson, and Ido M. Rosen. Multiple randomization designs. *CoRR*, abs/2112.13495, 2021.

7 Joeran Beel, Marcel Genzmehr, Stefan Langer, Andreas Nürnberger, and Bela Gipp. A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*, RepSys '13, page 7–14, New York, NY, USA, 2013. Association for Computing Machinery.

8 Joeran Beel and Stefan Langer. A comparison of offline evaluations, online evaluations, and user studies in the context of research-paper recommender systems. In Sarantos Kapidakis, Cezary Mazurek, and Marcin Werla, editors, *Proceedings of the 19th International Conference on Theory and Practice of Digital Libraries (TPDL)*, volume 9316 of *Lecture Notes in Computer Science*, pages 153–168, 2015.

9 Francois Belletti, Karthik Lakshmanan, Walid Krichene, Yi-Fan Chen, and John R. Anderson. Scalable realistic recommendation datasets through fractal expansions. *CoRR*, abs/1901.08910, 2019.

10 Alejandro Bellogín, Pablo Castells, and Iván Cantador. Precision-oriented evaluation of recommender systems: an algorithmic comparison. In Bamshad Mobasher, Robin D. Burke, Dietmar Jannach, and Gediminas Adomavicius, editors, *Proceedings of the 2011 ACM Conference on Recommender Systems, RecSys 2011, Chicago, IL, USA, October 23-27, 2011*, pages 333–336. ACM, 2011.

11 Alejandro Bellogín, Pablo Castells, and Iván Cantador. Statistical biases in information retrieval metrics for recommender systems. *Inf. Retr. J.*, 20(6):606–634, 2017.

12 Lonni Besançon and Pierre Dragicevic. The continued prevalence of dichotomous inferences at chi. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2019.

13 Tanja Blascheck, Lonni Besançon, Anastasia Bezerianos, Bongshin Lee, and Petra Isenberg. Glanceable visualization: Studies of data comparison performance on smartwatches. *IEEE transactions on visualization and computer graphics*, 25(1):630–640, 2018.

14 Jesús Bobadilla, Abraham Gutiérrez, Raciel Yera, and Luis Martínez. Creating synthetic datasets for collaborative filtering recommender systems using generative adversarial networks. *Knowl. Based Syst.*, 280:111016, 2023.

15 John S. Breese, David Heckerman, and Carl Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, page 43–52, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

16 Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In Mark Sanderson, Kalervo Järvelin, James Allan, and Peter Bruza, editors, *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004*, pages 25–32. ACM, 2004.

17 Rocío Cañamares and Pablo Castells. Should I follow the crowd?: A probabilistic analysis of the effectiveness of popularity in recommender systems. In Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz, editors, *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 415–424. ACM, 2018.

18 Rocío Cañamares and Pablo Castells. On target item sampling in offline recommender system evaluation. In Rodrygo L. T. Santos, Leandro Balby Marinho, Elizabeth M. Daly, Li Chen, Kim Falk, Noam Koenigstein, and Edleno Silva de Moura, editors, *RecSys 2020: Fourteenth ACM Conference on Recommender Systems, Virtual Event, Brazil, September 22-26, 2020*, pages 259–268. ACM, 2020.

19 Rocío Cañamares, Pablo Castells, and Alistair Moffat. Offline evaluation options for recommender systems. *Inf. Retr. J.*, 23(4):387–410, 2020.

20 Ozan Candogan, Chen Chen, and Rad Niazadeh. Correlated cluster-based randomized experiments: Robust variance minimization. *Management Science*, 2023.

21 Ben Carterette. Statistical significance testing in information retrieval: Theory and practice. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1387–1389, 2017.

22 Pablo Castells and Alistair Moffat. Offline recommender system evaluation: Challenges and new directions. *AI Mag.*, 43(2):225–238, 2022.

23 Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 621–630, 2009.

24 Jin Yao Chin, Yile Chen, and Gao Cong. The datasets dilemma: How much do we really know about recommendation datasets? In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 141–149, New York, NY, USA, 2022. Association for Computing Machinery.

25 Yashar Deldjoo, Tommaso Di Noia, Eugenio Di Sciascio, and Felice Antonio Merra. How dataset characteristics affect the robustness of collaborative recommendation models. In Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 951–960. ACM, 2020.

26 Pierre Dragicevic. Fair statistical communication in hci. *Modern statistical methods for HCI*, pages 291–330, 2016.

27 Michael D. Ekstrand, Ben Carterette, and Fernando Diaz. Distributionally-informed recommender system evaluation. *ACM Trans. Recomm. Syst.*, 2(1), mar 2024.

28 Bradley J. Erickson and Felipe Kitamura. Magician's corner: 9. performance metrics for machine learning models. *Radiology: Artificial Intelligence*, 3(3):e200126, 2021.

29 Ronald Aylmer Fisher. *Statistical methods for research workers.* Oliver and Boyd, Edinburgh, UK, 1st ed edition, 1925.

30 Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer, 1970.

31 Shir Frumerman, Guy Shani, Bracha Shapira, and Oren Sar Shalom. Are all rejected recommendations equally bad?: Towards analysing rejected recommendations. In George Angelos Papadopoulos, George Samaras, Stephan Weibelzahl, Dietmar Jannach, and Olga C. Santos, editors, *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2019, Larnaca, Cyprus, June 9-12, 2019*, pages 157–165. ACM, 2019.

32 Norbert Fuhr. Some common mistakes in ir evaluation, and how they can be avoided. *SIGIR Forum*, 51(3):32–41, feb 2018.

33 Florent Garcin, Boi Faltings, Olivier Donatsch, Ayar Alazzawi, Christophe Bruttin, and Amr Huber. Offline and online evaluation of news recommender systems at swissinfo.ch. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, page 169–176, New York, NY, USA, 2014. Association for Computing Machinery.

34 Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. Offline a/b testing for recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 198–206, New York, NY, USA, 2018. Association for Computing Machinery.

35 Alois Gruson, Praveen Chandar, Christophe Charbuillet, James McInerney, Samantha Hansen, Damien Tardieu, and Ben Carterette. Offline evaluation to make decisions about playlistrecommendation algorithms. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 420–428, New York, NY, USA, 2019. Association for Computing Machinery.

36 Asela Gunawardana, Guy Shani, and Sivan Yogev. Evaluating recommender systems. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 547–601. Springer US, 2022.

37 Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning.* Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

38 Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, jan 2004.

39 Balázs Hidasi and Ádám Tibor Czapp. Widespread flaws in offline evaluation of recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, page 848–855, New York, NY, USA, 2023. Association for Computing Machinery.

40 Katja Hofmann, Lihong Li, Filip Radlinski, et al. Online evaluation for information retrieval. *Foundations and Trends® in Information Retrieval*, 10(1):1–117, 2016.

41 Eugene Ie, Chih-Wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. RecSim: A configurable simulation platform for recommender systems. *CoRR*, abs/1909.04847, 2019.

42 Ngozi Ihemelandu and Michael D. Ekstrand. Candidate set sampling for evaluating Top-N recommendation. In *IEEE International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2023, Venice, Italy, October 26-29, 2023*, pages 88–94. IEEE, 2023.

43 Ngozi Ihemelandu and Michael D. Ekstrand. Inference at scale: Significance testing for large search and recommendation experiments. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2087–2091, New York, NY, USA, 2023. Association for Computing Machinery.

44 Ngozi Ihemelandu and Michael D. Ekstrand. Multiple testing for IR and recommendation system experiments. In Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis, editors, *Advances in Information Retrieval*, pages 449–457, Cham, 2024. Springer Nature Switzerland.

45 Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. *Evaluating recommender systems*, page 166–188. Cambridge University Press, 2010.

46 Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, oct 2002.

47 Olivier Jeunen. Revisiting offline evaluation for implicit-feedback recommender systems. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, page 596–600, New York, NY, USA, 2019. Association for Computing Machinery.

48 Olivier Jeunen. A common misassumption in online experiments with machine learning models. *SIGIR Forum*, 57(1), dec 2023.

49 Olivier Jeunen. A probabilistic position bias model for short-video recommendation feeds. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, page 675–681, New York, NY, USA, 2023. Association for Computing Machinery.

50 Olivier Jeunen and Ben London. Offline recommender system evaluation under unobserved confounding. In *RecSys 2023 Workshop: CONSEQUENCES – Causality, Counterfactuals and Sequential Decision-Making*, 2023.

51 Olivier Jeunen, Ivan Potapov, and Aleksei Ustimenko. On (normalised) discounted cumulative gain as an off-policy evaluation metric for top-n recommendation. In *Proceedings of the 30th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '24, New York, NY, USA, 2024. Association for Computing Machinery.

52 Olivier Jeunen, Koen Verstrepen, and Bart Goethals. Fair offline evaluation methodologies for implicit-feedback recommender systems with mnar data. In *Proceedings of the REVEAL Workshop on Offline Evaluation for Recommender Systems (RecSys '18)*, October 2018.

53 Yitong Ji, Aixin Sun, Jie Zhang, and Chenliang Li. A critical study on data leakage in recommender system offline evaluation. *ACM Trans. Inf. Syst.*, 41(3), feb 2023.

54 Sergey Kolesnikov and Mikhail Andronov. CVTT: cross-validation through time. *CoRR*, abs/2205.05393, 2022.

55 Walid Krichene and Steffen Rendle. On sampled metrics for item recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1748–1757, New York, NY, USA, 2020. Association for Computing Machinery.

56 Dongwon Lee, Anandasivam Gopal, and Sung-Hyuk Park. Different but equal? A field experiment on the impact of recommendation systems on mobile and personal computer channels in retail. *Inf. Syst. Res.*, 31(3):892–912, 2020.

57 Dong Li, Ruoming Jin, Jing Gao, and Zhi Liu. On sampling top-k recommendation evaluation. In Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash, editors, *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 2114–2124. ACM, 2020.

58 Xitong Li, Jörn Grahl, and Oliver Hinz. How do recommender systems lead to consumer purchases? A causal mediation analysis of a field experiment. *Inf. Syst. Res.*, 33(2):620–637, 2022.

59 Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. Variational autoencoders for collaborative filtering. In Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis, editors, *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, pages 689–698. ACM, 2018.

60 Yunji Liang, Lei Liu, Luwen Huangfu, Zhu Wang, and Bin Guo. Deepapp: characterizing dynamic user interests for mobile application recommendation. *World Wide Web (WWW)*, 26(5):2623–2645, 2023.

61 Benjamin M. Marlin and Richard S. Zemel. Collaborative prediction and ranking with non-random missing data. In Lawrence D. Bergman, Alexander Tuzhilin, Robin D. Burke, Alexander Felfernig, and Lars Schmidt-Thieme, editors, *Proceedings of the 2009 ACM Conference on Recommender Systems, RecSys 2009, New York, NY, USA, October 23-25, 2009*, pages 5–12. ACM, 2009.

62 Donald Metzler and Oren Kurland. Experimental methods for information retrieval. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1185–1186, 2012.

63 Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1):1–27, 2008.

64 Denis Parra and Shaghayegh Sahebi. Recommender systems: Sources of knowledge and evaluation metrics. In *Advanced Techniques in Web Intelligence-2: Web User Browsing Behaviour and Preference Analysis*, pages 149–175. Springer, 2013.

65 Oona Rainio, Jarmo Teuho, and Riku Klén. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1):6086, 2024.

66 Tomas Rehorek, Ondrej Biza, Radek Bartyzal, Pavel Kordik, Ivan Povalyev, and Ondrej Podsztavek. Comparing offline and online evaluation results of recommender systems. In *ACM Conference on Recommender Systems*, REVEAL Workshop, 2018.

67 Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: bayesian personalized ranking from implicit feedback. In Jeff A. Bilmes and Andrew Y. Ng, editors, *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*, pages 452–461. AUAI Press, 2009.

68 Allen L. Schirm Ronald L. Wasserstein and Nicole A. Lazar. Moving to a world beyond "p<0.05". *The American Statistician*, 73(sup1):1–19, 2019.

69 Marco Rossetti, Fabio Stella, and Markus Zanker. Contrasting offline and online results when evaluating recommendation algorithms. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, page 31–34, New York, NY, USA, 2016. Association for Computing Machinery.

70 Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

71 Alan Said and Alejandro Bellogín. Comparative recommender system evaluation: Benchmarking recommendation frameworks. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 129–136, 2014.

72 Yuta Saito and Thorsten Joachims. Counterfactual learning and evaluation for recommender systems: Foundations, implementations, and recent advances. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, page 828–830, New York, NY, USA, 2021. Association for Computing Machinery.

73 Tetsuya Sakai. Statistical reform in information retrieval? *ACM SIGIR Forum*, 48(1):3–12, 2014.

74 Tetsuya Sakai. Laboratory experiments in information retrieval. *The information retrieval series*, 40:4, 2018.

75 Tetsuya Sakai. On Fuhr's guideline for IR evaluation. *SIGIR Forum*, 54(1), feb 2021.

76 Pablo Sánchez and Alejandro Bellogín. Attribute-based evaluation for recommender systems: incorporating user and item attributes in evaluation metrics. In Toine Bogers, Alan Said, Peter Brusilovsky, and Domonkos Tikk, editors, *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019*, pages 378–382. ACM, 2019.

77 Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press, Cambridge, 2008.

78 Jing-Cheng Shi, Yang Yu, Qing Da, Shi-Yong Chen, and Anxiang Zeng. Virtual-taobao: Virtualizing real-world online retail environment for reinforcement learning. *CoRR*, abs/1805.10000, 2018.

79 Mark D. Smucker, James Allan, and Ben Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, page 623–632, New York, NY, USA, 2007. Association for Computing Machinery.

80 Harald Steck. Training and testing of recommender systems on data missing not at random. In Bharat Rao, Balaji Krishnapuram, Andrew Tomkins, and Qiang Yang, editors, *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, pages 713–722. ACM, 2010.

81 Harald Steck. Item popularity and recommendation accuracy. In Bamshad Mobasher, Robin D. Burke, Dietmar Jannach, and Gediminas Adomavicius, editors, *Proceedings of the 2011 ACM Conference on Recommender Systems, RecSys 2011, Chicago, IL, USA, October 23-27, 2011*, pages 125–132. ACM, 2011.

82 Yi Su, Xiangyu Wang, Elaine Ya Le, Liang Liu, Yuening Li, Haokai Lu, Benjamin Lipshitz, Sriraj Badam, Lukasz Heldt, Shuchao Bi, Ed H. Chi, Cristos Goodrow, Su-Lin Wu, Lexi Baugher, and Minmin Chen. Long-term value of exploration: Measurements, findings and algorithms. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 636–644, New York, NY, USA, 2024. Association for Computing Machinery.

83 Zhu Sun, Hui Fang, Jie Yang, Xinghua Qu, Hongyang Liu, Di Yu, Yew-Soon Ong, and Jie Zhang. DaisyRec 2.0: Benchmarking recommendation for rigorous evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8206–8226, 2023.

84 Lukas Svicarovic, Denis Parra, and María Jesús Lobo. Evaluating Interactive Comparison Techniques in a Multiclass Density Map for Visual Crime Analytics. In Marco Agus, Christoph Garth, and Andreas Kerren, editors, *EuroVis 2021 – Short Papers*. The Eurographics Association, 2021.

85 Ding Tong, Qifeng Qiao, Ting-Po Lee, James McInerney, and Justin Basilico. Navigating the feedback loop in recommender systems: Insights and strategies from industry practice.

In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, page 1058–1061, New York, NY, USA, 2023. Association for Computing Machinery.

86 Julián Urbano, Harlley Lima, and Alan Hanjalic. Statistical significance testing in information retrieval: An empirical analysis of type i, type ii and type iii errors. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 505–514, New York, NY, USA, 2019. Association for Computing Machinery.

87 Michalis Vafopoulos and Michael Oikonomou. Recommendation systems: Bridging technical aspects with marketing implications. In Ioannis Anagnostopoulos, Mária Bieliková, Phivos Mylonas, and Nicolas Tsapatsoulis, editors, *Semantic Hyper/Multimedia Adaptation – Schemes and Applications*, volume 418 of *Studies in Computational Intelligence*, pages 155–180. Springer, 2013.

88 Daniel Valcarce, Alejandro Bellogín, Javier Parapar, and Pablo Castells. On the robustness and discriminative power of information retrieval metrics for top-n recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, page 260–268, New York, NY, USA, 2018. Association for Computing Machinery.

89 Daniel Valcarce, Alejandro Bellogín, Javier Parapar, and Pablo Castells. Assessing ranking metrics in top-n recommendation. *Inf. Retr. J.*, 23(4):411–448, 2020.

90 Flavian Vasile, David Rohde, Olivier Jeunen, and Amine Benhalloum. A gentle introduction to recommendation as counterfactual policy learning. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, page 392–393, New York, NY, USA, 2020. Association for Computing Machinery.

91 Koen Verstrepen, Kanishka Bhaduriy, Boris Cule, and Bart Goethals. Collaborative filtering for binary, positiveonly data. *SIGKDD Explor. Newsl.*, 19(1):1–21, sep 2017.

92 Ellen M. Voorhees. Topic set size redux. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 806–807, New York, NY, USA, 2009. Association for Computing Machinery.

93 Zhiqiang Wan, Yazhou Zhang, and Haibo He. Variational autoencoder based synthetic data generation for imbalanced learning. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7, 2017.

94 Kai Wang, Zhene Zou, Minghao Zhao, Qilin Deng, Yue Shang, Yile Liang, Runze Wu, Xudong Shen, Tangjie Lyu, and Changjie Fan. Rl4rs: A real-world dataset for reinforcement learning based recommender system. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2935–2944, New York, NY, USA, 2023. Association for Computing Machinery.

95 Emine Yilmaz and Javed A. Aslam. Estimating average precision with incomplete and imperfect judgments. In Philip S. Yu, Vassilis J. Tsotras, Edward A. Fox, and Bing Liu, editors, *Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management, Arlington, Virginia, USA, November 6-11, 2006*, pages 102–111. ACM, 2006.

96 Kesen Zhao, Shuchang Liu, Qingpeng Cai, Xiangyu Zhao, Ziru Liu, Dong Zheng, Peng Jiang, and Kun Gai. Kuaisim: A comprehensive simulator for recommender systems. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

4.2 Fairness Evaluation

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

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

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

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

4.2.1 Landscape

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

For "Who"?

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

For Who	On What Basis	How it harms	Consequences

• Item
• Item Entities
• Item Providers
• Consumers
	• Individual
	• Group
	  • What Groups?
	  • Which Attributes?
		Short term impact	Long term impact

Figure 1 Categorization of fairness factors.

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

On "What" basis?

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

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

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

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

"How"?

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

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

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

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

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

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

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

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

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

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

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

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

On "What" Scale?

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

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

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

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

4.2.2 Examples / Use cases

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

Research paper recommender system/search engine

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

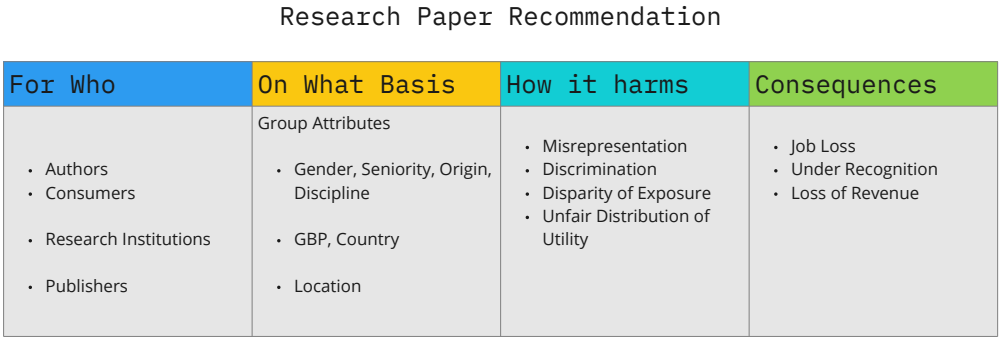

Figure 2 Identifying the key points of fairness in research paper recommender systems.

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

Examples of fairness concerns for this domain include:

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

E-commerce

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

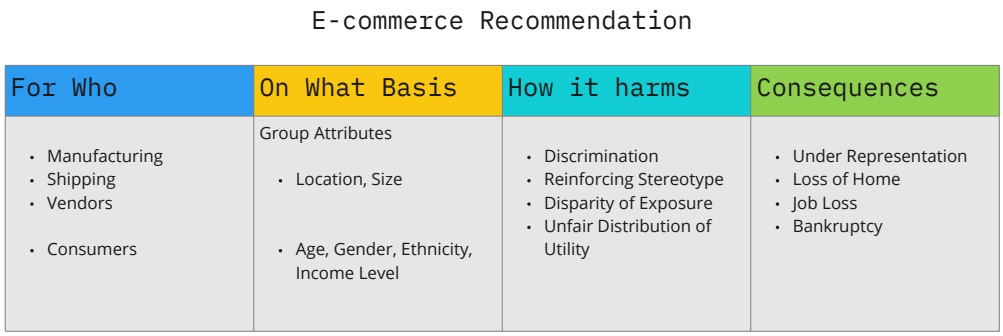

Figure 3 Identifying the key points of fairness in e-commerce recommender systems.

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

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

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

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

4.2.3 Problem definition

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

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

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

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

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

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

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

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

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

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

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

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

4.2.4 Operationalization & Planning

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

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

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

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

4.2.4.1 Scope of measurement

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

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

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

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

4.2.4.2 Inputs to evaluation

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

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

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

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

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

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

4.2.4.3 Experiment design

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

4.2.4.4 Choosing measurements

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

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

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

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

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

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

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

4.2.4.5 Iterating on operationalization

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

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

4.2.5 Analysis & interpretation

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

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

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

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

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

4.2.6 Reporting & sharing

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

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

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

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

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

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

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

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

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

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

In summary:

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

4.2.7 Conclusion

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

17 Michael D Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. Fairness in information access systems. *Foundations and Trends® in Information Retrieval*, 16(1-2):1–177, 2022.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

39 Amifa Raj and Michael D Ekstrand. Measuring fairness in ranked results: An analytical and empirical comparison. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 726–736. ACM, July 2022.

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

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

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

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

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

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

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

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

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

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

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

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

4.3 Best-Practices for Offline Evaluations of Recommender Systems

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

4.3.1 Introduction

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

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

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

We chose here to focus on the scope of *offline evaluation*, identifying problems and best practices for this type of evaluation. While recommender systems are not only evaluated offline, this evaluation still represents a significant part of many recommender system papers. Furthermore, we limited the scope to only offline evaluation to provide concrete focused tools that can be implemented immediately. We believe the same ideas and goals that guided us throughout this report can later be extended to encompass other evaluation processes, such as user studies, A/B testing, and more.

The rest of this section is organized as follows. After discussing previous works in Section 4.3.2, we outline the main challenges regarding recommender systems evaluation concerning reproducibility and methodology in Section 4.3.3. Section 4.3.4 then contains specific guidelines in the form of key questions in this context to be answered by paper authors and/or reviewers when preparing or reviewing a paper. Furthermore, this section provides recommendations for editors and program chairs regarding the possible implementation of these measures and potential risks.

4.3.2 Related Work

While the recommender systems research area is increasing rapidly in terms of research publications, there are currently no clear, agreed-upon, and widely adopted guidelines for critical aspects of empirical evaluation. This section provides a brief overview of work in recommender systems that analyzes, reflects, and criticizes the literature concerning empirical evaluation. It links the problems in the recommender systems community to other fields that experience similar problems. Finally, it presents potential solutions based on communities that have undergone similar challenges and converged on a set of guidelines that the entire community follows.

Despite the continuously increasing popularity of the recommender systems research field and the demonstrated good performance of recent recommender systems, there is a notable lack of standardized criteria or methods for evaluating their performance. The work by [35] represents one recent example that highlights this as a problem.

Among the existing approaches towards some form of standardization, researchers and practitioners in the field have previously proposed using different evaluation frameworks. *Software frameworks* often implement a particular evaluation protocol and support specific metrics, promoting a set of standards. Examples of these are Elliot [1], LensKit [15], RecBole [45], and RecPack [27] to name a few. On the other hand, there are also *conceptual frameworks* such as FEVR [48] and the replicable recommendation process presented by [10]. However, as different frameworks and packages use different protocols for the various steps in the *preprocessing-recommendation-evaluation pipeline*, they remain somewhat limited in their capability to help the community converge on an agreed-upon set of guidelines.

Beyond the realm of recommender systems, research in machine learning, in general, has previously been criticized, specifically pointing out that the field is undergoing a "reproducibility and replication" crisis to the extent that parts of the community are suggesting that research results and claims cannot be taken at face value [22]. The field of information retrieval has seen similar experiences for extended periods of time suggesting, e.g., that reported improvements are not reflecting actual improvements [2], and optimization based solely on aggregated measurements can potentially lead to misleading and unreliable outcomes [46]. These insights are not unique to applied fields such as recommendation and retrieval.

In evidence synthesis[8], similar insights have been identified, specifically pointing out how the validity and reproducibility of meta-analyses are affected by poorly documented and biased data collection [25].

Turning our attention again to recommender systems, reproducibility research in this area has surfaced over the years as an increasingly more important aspect, specifically highlighted by influential papers. For example, [19, 18] point out that much of the improvements reported for certain algorithmic approaches were rather "phantom progress" than actual. Also, [71] show that design choices in implementing algorithms and evaluation methods in widely used software packages for recommendation lead to large differences in performance between frameworks even when using identical datasets, settings, and evaluation strategies. [6] showed that identical algorithms performed vastly differently on relatively similar news platforms. More recently, [38] identified that hyperparameter tuning (or lack thereof) can lead to inaccurate comparisons between introduced state of the art methods and widely-used baselines.

The research communities attempted to address these challenges through the frameworks discussed above and through initiatives that are supposed to help foster reproducibility and transparency in evaluation. Examples of these include the Reproducibility track introduced at ACM RecSys in 2020[9] with similar initiatives having been established at related conferences. Another example includes the submission type "Registered Reports" in ACM TORS[10].

In information retrieval, a recent example of guidelines addressing the above-mentioned issues includes a checklist to "strengthen an IR paper" [39], which was published for the SIGIR 2022 conference. The checklist is split into two parts. One part covers the presentation and writing of a manuscript. It contains seven bullet points, which are mostly framed at a high level ("The results are presented effectively in the appropriate format"). The second part addresses the experimental design with six bullet points, also at a high level ("The experimental results are reliable and generalizable"). Similarly, the SIGIR-AP conference provides its authors with guidelines [40]. However, these may appear to be rather short and high-level. For instance, the guidelines specify that "The experimental design and its scale [should be] appropriate to the problem". How an appropriate experimental design actually might look is not detailed.

Perhaps the most related work to ours is the best practice guidelines and checklists from the *machine learning* community. Premier machine-learning conferences such as NeurIPS introduced guidelines and checklists a few years ago [12, 28] and continue to use them today [31]. NeurIPS 2024 provides a 15-item checklist with guidelines [31]. The checklist is incorporated into the LaTeX paper submission (Fig. 4) template [30]. Authors must answer and submit the checklist along with their manuscript; otherwise, their submission will be desk-rejected. These questions relate to various aspects of the work, including the validity of claims, reproducibility, open access, and ethical considerations. Authors may answer with "yes", "no" or "n/a" and can provide one or two sentences of justification. The questions are more specific than those of the SIGIR conference. For instance, concerning experimental design, one question reads: *"Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?"*.

[8] Evidences synthesis *"refers to the process of bringing together information from a range of sources and disciplines to inform debates and decisions on specific issues."*, `https://royalsociety.org/news-resources/projects/evidence-synthesis/`

[9] `https://recsys.acm.org/recsys20/call/#content-tab-1-1-tab`

[10] `https://dl.acm.org/journal/tors/author-guidelines`

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [TODO]
   (b) Did you describe the limitations of your work? [TODO]
   (c) Did you discuss any potential negative societal impacts of your work? [TODO]
   (d) Did you read the ethics review guidelines and ensure that your paper conforms to them? `https://2022.automl.cc/ethics-accessibility/` [TODO]
2. If you ran experiments...
   (a) Did you use the same evaluation protocol for all methods being compared (e.g., same benchmarks, data (sub)sets, available resources)? [TODO]
   (b) Did you specify all the necessary details of your evaluation (e.g., data splits, pre-processing, search spaces, hyperparameter tuning)? [TODO]
   (c) Did you repeat your experiments (e.g., across multiple random seeds or splits) to account for the impact of randomness in your methods or data? [TODO]
   (d) Did you report the uncertainty of your results (e.g., the variance across random seeds or splits)? [TODO]

Figure 4 Screenshot of the AutoML conference submission checklist [3]. The list is published under CC-BY 4.0 license.

NeurIPS is experimenting with large language models to provide an assistant that supports authors with the checklist [32]. NeurIPS also "strongly encourages" [29] authors to submit their code and data and follow the guidelines set forth by the "Papers with Code" platform [44]. These code submission guidelines provide a code template for installing, training, and evaluating machine learning models and a template for downloading pre-trained models [44]. Besides NeurIPS, other conferences such as MICCAI adopted the code template as well [26]. Furthermore, the NeurIPS checklist [28] was also adopted by other conferences, including ICML [23].

To our knowledge, the most comprehensive list of guidelines is the AutoML conference submission checklist [3]. Like NeurIPS, the checklist is directly incorporated into the manuscript template, and authors must submit the checklist as an appendix. Answers to 28 questions have to be given, commonly with the options "yes", "no" and "n/a". A short justification or reference is required for every answer; see Fig. 4 for a screenshot. For instance, for the question "*Did you include the license for the code and dataset?*", an author may answer with "yes" and refer to details in the paper, e.g., by writing "*See Section 7 in the manuscript*". The questions used in the context of the AutoML conference are more specific than those of NeurIPS.

4.3.3 Addressed Problem Areas

We focus on two main areas that may hamper progress in our field [13]: (a) *reproducibility* and (b) problematic practices in terms of evaluation *methodology*.

4.3.3.1 Reproducibility – Purpose and Definition

Reproducibility refers to the ability to achieve the same findings as the original researchers utilizing existing data from a prior study [20]. In other words, reproducibility describes the minimum necessary information to re-implement, re-execute, repeat, and replicate experiments to verify the findings described in a scientific study [11, 20, 36].

While different and partially incompatible definitions of reproducibility and related concepts exist, in computer science, reproducibility often implies that experiments, including data processing steps, can be accurately repeated to produce the same results. This typically

involves making code and data publicly accessible and providing detailed documentation of computational methods and algorithms. This section adopts this notion as a working definition of reproducibility.

Generally, reproducibility in recommender systems research is essential, as it allows researchers to ensure that others can (a) verify previously published results and (b) make sure that their algorithmic contributions truly help to advance state of the art.

To shed light on our concerns, we list several issues often observed in papers concerning reproducibility. First and foremost, while we observed a positive development over the past decade, many researchers and practitioners still do not publish their algorithmic implementation. While Intellectual Property (IP) rights issues may pose a challenge in some cases, it is still important that the implementations are made available to others.

However, publishing the proposed model or method implementation is not enough. For reproducibility, the code for the entire evaluation pipeline is required, starting with loading the data and ending with results. When doing so, one should focus on the proposed model and the implementation and training of the used baselines. Specifically, tuning the baselines, i.e., how the hyperparameters were determined, should be made public.

4.3.3.2 Evaluation Methodology – Purpose and Definition

Scientific evaluation is a systematic approach used to assess the validity of a hypothesis. Evaluation methodology outlines the proper conduct of scientific evaluation [14]. The evaluation methodology is typically characterized by a concrete set of steps to ensure rigorous scientific standards set by the community. It is driven by the underlying hypothesis or research questions to ensure that assessments and conclusions drawn from the empirical results are scientifically sound [34]. For a given research problem, researchers commonly make decisions in their evaluation methodology that are inspired and justified by previous research.

The evaluation methodology in offline experiments for recommender system research typically details the collection and preprocessing of data, the chosen baselines, the learning and optimization strategy, the metrics used to compare methods, and the method used to analyze the results.

The evaluation methodology is critical in scientific work because it allows researchers to create an evaluation process that others can validate and reuse for similar research. An extensive description of the evaluation methodology is also essential for the peer-reviewing process. It allows reviewers and other researchers to critically assess the validity of the results obtained to confirm or refute a hypothesis. The peer-reviewing process should thus ensure that the evaluation methodology in a research paper is rigorous and follows the community's consensus.

In the following, we identify a few common problematic recommender systems evaluation methodology practices that our guidelines intend to remedy. First, we find that researchers often provide little or no justifications for certain choices of their research methodology. Researchers often justify certain decisions by arguing that *the decision is common practice* or that *another group of researchers did it*. Adopting methodological choices from previous works can be beneficial, making research more comparable. However, such a justification may often not be scientific or complete, e.g., when the *the common practice* had no proper justification either.

Second, we note that data leakage is a common problem in evaluation methods that may be unseen by just reading the evaluation methodology. Data leakage refers to using the *test* split or knowledge about the test split in the training process [21].

Finally, while hyperparameter optimization is a standard procedure in recommender systems research, the configurations are often not shared or are incomplete. However, it has been shown [38] that the performance difference between configurations can be significant and change the ranking of algorithms. This is especially problematic in research that claims to improve over the state-of-the-art but provides incomplete hyperparameter tuning specifications for baseline algorithms.

4.3.4 Proposed Measures

This section outlines two catalogs of questions regarding reproducibility and experimental choices. These questions should serve as a basis for implementing concrete measures for relevant publications outlets, e.g., author guidelines, author self-assessment forms, reviewer guidelines, or extended paper review forms. Further questions may be added depending on the chosen purpose and implementation, while others might be left out. If some of the questions are unanswered, the authors should have a good justification for why they are irrelevant to their research project. In any case, the paper should contain answers to all the relevant questions within the text below.

We formulate these lists as high-level questions, followed by a list of issues one must consider when answering these questions. We also make the questions available as a LaTeXtemplate [11].

4.3.4.1 Author and Reviewer Checklist: Reproducibility

We organize the proposed questions on reproducibility (and their corresponding explanations) in the following groups.

1. Code-related Aspects: Is the code of the full experimental pipeline publicly available?
   Sharing all artifacts needed or used to obtain the numerical results reported in a paper is essential for reproducibility. Appropriate documentation must also be provided so other researchers can re-execute the experiments. If possible, an execution-ready environment, e.g., in terms of a Docker container, should be made available.
   - **1.1.** Code of proposed algorithm/framework/method/model
   - **1.2.** Code of all baselines
   - **1.3.** Code for preprocessing and postprocessing
   - **1.4.** Code for hyperparameter tuning
   - **1.5.** Code for execution (training and testing)
   - **1.6.** Code for statistical analysis
   - **1.7.** Documentation and installation/execution instructions
2. Data-related Aspects: Is all relevant data publicly available?
   Reproducibility is only possible if the data used as input to the models and the results are publicly available. It may be insufficient to provide pointers only to previously published datasets, e.g., because preprocessing steps have been applied or publicly shared datasets are sometimes updated.
   - **2.1.** Original datasets
   - **2.2.** Preprocessed datasets
   - **2.3.** Train/validation/test splits
   - **2.4.** Results (outcomes of measurements)
   - **2.5.** Trained models

[11] `https://code.recommender-systems.com/Dagstuhl-24211-Checklist`

3. Configuration Aspects: Are all relevant configuration parameters reported?
   Besides code and data, the specifics of the execution of the experiment must be documented. This concerns how the models were tuned, the execution environment, and its configuration.
   - **3.1.** Hyperparameter search strategy, search space and search time for all models
   - **3.2.** Optimal hyperparameters per dataset and model
   - **3.3.** Train-test splitting configurations
   - **3.4.** Random seeds
   - **3.5.** Required external libraries and their versions
   - **3.6.** Used hardware (configuration)
4. Experiment specific aspects and other questions
   Depending on the specifics of the experiment, information about various other aspects should be provided. These questions should help better to gauge the level of reproducibility of the experiment. Further, these questions may serve as a place for researchers to justify certain technical choices.
   - **4.1.** Has an existing evaluation framework been used? If not, why not?
   - **4.2.** Is "one-click" reproducibility supported?
   - **4.3.** Are any instructions provided to reproduce (at least parts of) the experiment with limited hardware resources?
   - **4.4.** Is an expected runtime to reproduce the results provided?

4.3.4.2 Author and Reviewer Checklist: Methodology

For a mature research community, embracing the evaluation procedure used in previous papers can be considered good practice. However, we must acknowledge that several unjustified protocols, e.g., leave-one-out, have taken root in the recommender system community. Hence, justifying a research protocol only by saying it was used in previous papers is perhaps unreasonable.

1. Research Questions and Hypothesis: Are the research questions and hypotheses expressed clearly and matching the method and the results?
   The research question should guide the development of the evaluation process. Therefore, it should be clearly stated, and the authors' choices throughout the method should correspond with the research question and conclusions.
   - **1.1.** The research question is clearly stated.
   - **1.2.** The hypothesis is derived from the research question.
   - **1.3.** The experimental design is suited to address the stated research questions.
   - **1.4.** The conclusion is based on the research question and the experimental design.
2. Baselines: Are baselines selected and tuned to ensure appropriate comparisons?
   While one should always compare to the latest best method for the particular task, it is also important to compare against earlier and probably simpler baselines to show the advantage of using the new, more complicated method.
   - **2.1.** The chosen baselines are appropriate to the hypothesis and research question.
   - **2.2.** One of the baselines is successful, e.g., state-of-the-art, for the given task.
   - **2.3.** At least one simple baseline, e.g., kNN, popularity, or random, is included.
   - **2.4.** The baselines are tuned. One must invest sufficient effort in properly training the baselines.
   - **2.5.** There needs to be clarity about whether the baselines were rerun or whether the results were taken from a previous paper.

3. Evaluation Metrics: Is the chosen evaluation metric appropriate to answer the research question?
   Choosing the appropriate evaluation metric for the task is critical. Reporting a large number of unrelated metrics is not good practice.
   - **3.1.** The selected metrics are derived from the hypothesis, e.g., RMSE for rating prediction or precision@N for top-N recommendation.
   - **3.2.** The reported metrics are not redundant, e.g., RMSE and MAE or DCG and NDCG.
   - **3.3.** Tradeoffs between the metrics are explained and evaluated.
4. Data collection: Is the data collection process reasonable and well explained?
   This is appropriate when a new dataset is presented. This dataset may be collected from an already running system or using a particular user study.
   - **4.1.** The data collection process is clear.
   - **4.2.** The study participants' recruitment, introduction, and participation incentives are explained.
   - **4.3.** Biases that exist in the data or arise from the data collection process are explained.
   - **4.4.** The used datasets are publicly available.
5. Datasets: Are the chosen datasets appropriate for the task?
   In offline evaluation, choosing appropriate datasets is highly important. Using a diverse set of datasets supports claims for generalization. In cases where a particular domain is targeted, the datasets must be focused on the task.
   - **5.1.** The chosen datasets are appropriate to the task at hand.
   - **5.2.** It should be clear whether the datasets were chosen to demonstrate generalization.
   - **5.3.** In the generalization case, it is desirable to experiment with a sufficient number of datasets.
   - **5.4.** If showing the general applicability of a model is the goal, a diverse set of datasets is used.
   - **5.5.** The origins of public datasets are specified.
6. Data preprocessing: Is the data preprocessing well justified and explained?
   It is often the case that researchers preprocess, prune, and filter the original dataset before training and testing. In general, preprocessing should be discouraged, especially the dataset's filtering and pruning. Such preprocessing should be kept to a minimum and should be well explained.
   - **6.1.** If users or items were pruned from the dataset, the pruning is well justified.
   - **6.2.** When pruning is done because the evaluated method works better on a subset of the data, this is made clear.
   - **6.3.** : This process is clearly explained and justified if the data was converted, e.g., from numeric ratings to binary like/dislike.
7. Data-splitting: Does the train-test split fit the structure of the dataset and the task?
   Most machine learning methods require a training phase. It is, hence, standard practice to split the data into training and test sets, where the test set is used only once to evaluate the algorithm once the training phase is done. The train-test split is designed to simulate the behavior at run time, when the system is aware of all information to date and must make future recommendations. Hence, the split procedure should correspond to the task at hand.
   - **7.1.** Typically, user-item interactions are split on time, where the training data contains the earlier interactions, and the test data contains the newer ones. When other types of splits are used, this is justified.
   - **7.2.** All algorithms are run on the same train-test split.
   - **7.3.** Cross-validation is applied when possible.

8. Hyperparameter Optimization (HPO): Is the hyperparameter optimization procedure justified and appropriate for the task?
   For many machine learning methods, it is well known that HPO is a critical factor for performance. ML algorithms may underperform significantly when their parameters are not tuned to the dataset. Using an organized HPO process for all evaluated algorithms is highly important.
   - **8.1.** The optimization strategy is clearly stated.
   - **8.2.** The hyperparameter configuration space (parameter range) is sufficiently large and clearly defined.
   - **8.3.** The optimization time or number of tested configurations is clearly stated.
   - **8.4.** It is stated in case some algorithms are optimized differently.
9. Experiment execution: Was the experiment executed such that the comparison results are fair and statistically sound?
   When running the experiments, all algorithms should receive equal treatment. Statistical significance should be computed to test the likelihood that the observed differences between the algorithms are real.
   - **9.1.** The boundaries between train and test data are respected (i.e., test data not used for checking convergence).
   - **9.2.** There is equal treatment of all compared algorithms (with respect, e.g., to HPO, runtime, hardware).
   - **9.3.** The statistical significance testing method is appropriate for the task.
   - **9.4.** The p-values are properly computed and reported.
   - **9.5.** Confidence intervals are provided whenever possible.
   - **9.6.** The hardware used in the experiment (e.g., memory, processor speed, GPU) is properly described.
10. Sensitivity analysis: Did the authors conduct and report a sensitivity analysis concerning the method parameters and the dataset properties?
    Many algorithms have some parameters that must be tuned. It is important to analyze how different values for these parameters influence the performance. In many cases, an algorithm may also be sensitive to the dataset's properties (e.g., sparsity).
    - **10.1.** The method is executed with different parameter values.
    - **10.2.** The values of all parameters are fixed except for the tested one.
    - **10.3.** The effect of the parameters on the method is reported and discussed.
    - **10.4.** If there are trade-offs between the parameters, they are made clear.
    - **10.5.** Sensitivity to dataset parameters is done similarly to the method parameters.

4.3.4.3 Practical Implementation Suggestions

In this section, we provide several concrete suggestions that could be immediately implemented by the ACM RecSys program chairs, the ACM TORS[12] editorial board, and the chairs and editors of related publications venues.

Author Checklist. We suggest choosing some of the questions above to create an author checklist that must be submitted alongside the paper. The checklist could be similar to the NeurIPS [31] and AutoML checklists [3], where authors answer the questions with [Yes], [No] or [N/A], and provide an explanation. For example, a question may be "Have you used a diverse set of datasets?" and the author may answer "No, because this paper is about a particular recommendation domain and does not naturally generalize to other problems."

[12] `https://dl.acm.org/journal/tors/`

In some conferences and journals, when a paper is accepted for publication, this checklist is published alongside the paper as an appendix, allowing readers to understand the rationale behind the authors' choices when designing their experiments. Both RecSys and TORS can embrace this suggestion. We believe that once authors know that they must explain their choices, they will make more informed choices.

Reviewer questions. In most conferences, the reviewers must answer closed questions alongside their free-text review. For example, reviewers are often asked to rate the novelty or significance of the work. We suggest adding several such closed questions concerning the evaluation procedure.

For example, such a question may be: "Is the choice of baseline methods appropriate (e.g., did they compare to basic methods, did they compare against recent methods, did they invest a reasonable effort into optimizing the baselines)?". The answer can be numeric or on a scale from 1-5, allowing for some flexibility.

We believe that once the reviewers are forced to answer these questions, this may also reflect on their final acceptance score. For example, if a reviewer sees that a paper has followed the best practices implied by these questions, it will strengthen the paper's chance of getting accepted, and vice versa. There may certainly be concerns that adding too many questions may cause the reviewers to avoid writing detailed reviews. Hence, we suggest restricting the number of questions to three to five items related to the empirical evaluation.

Outstanding Methodology Research Papers. ACM RecSys, like other conferences, has the best paper award and best student paper award. We suggest adding, alongside these awards, an "outstanding evaluation practices paper" award, whose evaluation would be centred around the questions detailed above. Similarly, journals like ACM TORS could implement such an award.

This award would be geared towards papers that conducted a particular empirical evaluation that goes beyond the standard best practices. For example, these papers may have an impressive comparison with many baselines over many datasets or suggest a new, well-motivated experimental design. As with other awards, the reviewers can propose candidate papers for this category, and then a committee will choose outstanding papers. The award will be given at the conference award ceremony.

Best-Practice Methodology Paper Track. We suggest adding a new track to ACM RecSys, inviting authors to submit papers that describe, rather than a new algorithmic innovation or a new domain for recommender system, a description of an evaluation methodology. For example, such a paper can review best practices in a particular sub-area, such as how one should evaluate multistakeholder recommendation algorithms. Alternatively, such a paper can suggest a novel method for conducting a particular experiment in a specific domain.

4.3.5 Conclusion

Today, researchers use a variety of ways to evaluate recommendation algorithms, making it difficult to assess the progress made in our field. This problem is aggravated by a certain lack of reproducibility of published research. One way to address this problem is to provide detailed guidelines for authors and reviewers regarding questions of methodology and reproducibility. We find that such guidelines are becoming increasingly used by conferences in the broader field of machine learning.

In this section, we propose a specific set of guidelines for recommender systems research, which conference program chairs and journal editors can rely on when implementing measures to improve the scientific rigor of research published in our field. In future works, we believe there is great potential in looking into domains beyond computer science to learn how guidelines are designed to be effective, e.g., in the medical domain [5].

References

1 Vito Walter Anelli, Alejandro Bellogin, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco Maria Donini, and Tommaso Di Noia. Elliot: A comprehensive and rigorous framework for reproducible recommender systems evaluation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2405–2414, New York, NY, USA, 2021. Association for Computing Machinery.

2 Timothy G. Armstrong, Alistair Moffat, William Webber, and Justin Zobel. Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, page 601–610, New York, NY, USA, 2009. Association for Computing Machinery.

3 AutoML. Automl author instructions. *https://github.com/automl-conf/LatexTemplate/blob/main/instructions.pdf*, 2024.

4 Joeran Beel. Releasing recsys research code. *https://github.com/ISG-Siegen/Releasing-RecSys-Research-Code*, 2023.

5 Joeran Beel. Proposal for evidence-based best-practices forrecommender-systems evaluation. In Christine Bauer, Alan Said, and Eva Zangerle, editors, *Evaluation Perspectives of Recommender Systems: Driving Research and Education*, volume 24211 of *Report from Dagstuhl Seminar*, 2024.

6 Joeran Beel, Corinna Breitinger, Stefan Langer, Andreas Lommatzsch, and Bela Gipp. Towards reproducibility in recommender-systems research. *User Modeling and User-Adapted Interaction (UMUAI)*, 26(1):69–101, 2016.

7 Joeran Beel, Timo Breuer, Anita Crescenzi, Norbert Fuhr, and Meije Li. Results-blind reviewing. *Dagstuhl Reports*, 13(1):68–154, 2023.

8 Joeran Beel and Victor Brunel. Data pruning in recommender systems research: Best-practice or malpractice? In *13th ACM Conference on Recommender Systems (RecSys)*, volume 2431, pages 26–30. CEUR-WS, 2019.

9 Alejandro Bellogín and Alan Said. Offline and online evaluation of recommendations. In Shlomo Berkovsky, Iván Cantador, and Domonkos Tikk, editors, *Collaborative Recommendations – Algorithms, Practical Challenges and Applications*, pages 295–328. WorldScientific, 2018.

10 Alejandro Bellogín and Alan Said. Improving accountability in recommender systems research through reproducibility. *User Model. User Adapt. Interact.*, 31(5):941–977, 2021.

11 Fabien C. Y. Benureau and Nicolas P. Rougier. Re-run, repeat, reproduce, reuse, replicate: Transforming code into scientific contributions. *Frontiers in Neuroinformatics*, 11, 2018.

12 Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. Introducing the neurips 2021 paper checklist. *https://neuripsconf.medium.com/introducing-the-neurips-2021-paper-checklist-3220d6df500b*, 2021.

13 Paolo Cremonesi and Dietmar Jannach. Progress in recommender systems research: Crisis? what crisis? *AI Magazine*, 42(3):43–54, 2021.

14 E Jane Davidson. *Evaluation methodology basics: The nuts and bolts of sound evaluation*. Sage, 2005.

15 Michael D. Ekstrand. LensKit for Python: Next-generation software for recommender systems experiments. In *Proceedings of the 29th ACM International Conference on Inform-*

ation & Knowledge Management, CIKM '20, page 2999–3006, New York, NY, USA, 2020. Association for Computing Machinery.

16 Michael D. Ekstrand. seedbank: easy management of seeds across rngs. May 2024.

17 Michael D. Ekstrand, Michael Ludwig, Joseph A. Konstan, and John T. Riedl. Rethinking the recommender research ecosystem: reproducibility, openness, and LensKit. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, page 133–140, New York, NY, USA, 2011. Association for Computing Machinery.

18 Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. A troubling analysis of reproducibility and progress in recommender systems research. *ACM Transactions on Information Systems*, 39(2), 2021.

19 Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, page 101–109, New York, NY, USA, 2019. Association for Computing Machinery.

20 National Science Foundation. Companion guidelines on replication & reproducibility in education research, 2018.

21 Awni Hannun, Chuan Guo, and Laurens van der Maaten. Measuring data leakage in machine-learning models with fisher information. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 760–770. PMLR, 27–30 Jul 2021.

22 Jessica Hullman, Sayash Kapoor, Priyanka Nanayakkara, Andrew Gelman, and Arvind Narayanan. The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 335–348, New York, NY, USA, 2022. Association for Computing Machinery.

23 ICML. Icml 2023 paper guidelines. *https://icml.cc/Conferences/2023/PaperGuidelines*, 2023.

24 Joseph A. Konstan and Gediminas Adomavicius. Toward identification and adoption of best practices in algorithmic recommender systems research. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*, RepSys '13, page 23–28, New York, NY, USA, 2013. Association for Computing Machinery.

25 Arielle Marks-Anglin and Yong Chen. A historical review of publication bias. *Research Synthesis Methods*, 11(6):725–742, 2020.

26 MICCAI. Miccai-reproducibility-checklist. *https://github.com/JunMa11/MICCAI-Reproducibility-Checklist/blob/main/README.md*, 2024.

27 Lien Michiels, Robin Verachtert, and Bart Goethals. Recpack: An(other) experimentation toolkit for top-n recommendation using implicit feedback data. In *Proceedings of the 16th ACM Conference on Recommender Systems*, RecSys '22, page 648–651, New York, NY, USA, 2022. Association for Computing Machinery.

28 NeurIPS. Neurips 2022 paper checklist guidelines. *https://neurips.cc/Conferences/2022/PaperInformation/PaperChecklist*, 2022.

29 NeurIPS. Neurips code and data submission guidelines. *https://neurips.cc/public/guides/CodeSubmissionPolicy*, 2024.

30 NeurIPS. Neurips latex style file. *https://media.neurips.cc/Conferences/NeurIPS2024/Styles.zip*, 2024.

31 NeurIPS. Neurips paper checklist guidelines. *https://neurips.cc/public/guides/PaperChecklist*, 2024.

32 NeurIPS. Soliciting participants for the neurips 2024 checklist assistant study. *https://blog.neurips.cc/2024/05/07/soliciting-participants-for-the-neurips-2024-checklist-assistant-study/*, 2024.

33 Steffen Rendle, Walid Krichene, Li Zhang, and Yehuda Koren. Revisiting the performance of ials on item recommendation benchmarks. In Jennifer Golbeck, F. Maxwell Harper, Vanessa Murdock, Michael D. Ekstrand, Bracha Shapira, Justin Basilico, Keld T. Lundgaard, and Even Oldridge, editors, *RecSys '22: Sixteenth ACM Conference on Recommender Systems, Seattle, WA, USA, September 18 – 23, 2022*, pages 427–435. ACM, 2022.

34 Peter H Rossi, Mark W Lipsey, and Gary T Henry. *Evaluation: A systematic approach*. Sage publications, 2018.

35 Deepjyoti Roy and Mala Dutta. A systematic review and research perspective on recommender systems. *J. Big Data*, 9(1):59, 2022.

36 Alan Said and Alejandro Bellogín. Replicable evaluation of recommender systems. In Hannes Werthner, Markus Zanker, Jennifer Golbeck, and Giovanni Semeraro, editors, *Proceedings of the 9th ACM Conference on Recommender Systems, RecSys 2015, Vienna, Austria, September 16-20, 2015*, pages 363–364. ACM, 2015.

37 Teresa Scheidt and Joeran Beel. Time-dependent evaluation of recommender systems. In *Perspectives on the Evaluation of Recommender Systems Workshop, ACM RecSys Conference*, 2021.

38 Faisal Shehzad and Dietmar Jannach. Everyone's a winner! on hyperparameter tuning of recommendation models. In *17th ACM Conference on Recommender Systems*, 2023.

39 SIGIR. Checklist to strengthen an ir paper. *https://sigir.org/sigir2022/checklist-to-strengthen-an-ir-paper/*, 2022.

40 SIGIR-AP. Guidance for authors. *https://www.sigir-ap.org/sigir-ap-2023/guidance-for-authors/*, 2023.

41 Zhu Sun, Di Yu, Hui Fang, Jie Yang, Xinghua Qu, Jie Zhang, and Cong Geng. Are we evaluating rigorously? benchmarking recommendation for reproducible evaluation and fair comparison. In *Proceedings of the 14th ACM Conference on Recommender Systems*, RecSys '20, page 23–32, New York, NY, USA, 2020. Association for Computing Machinery.

42 Tobias Vente, Michael Ekstrand, and Joeran Beel. Introducing LensKit-Auto, an experimental automated recommender system (autorecsys) toolkit. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1212–1216, 2023.

43 Lukas Wegmeth, Tobias Vente, Lennart Purucker, and Joeran Beel. The effect of random seeds for data splitting on recommendation accuracy. In *Proceedings of the 3rd Perspectives on the Evaluation of Recommender Systems Workshop*, 2023.

44 Papers with code. Code template. *https://github.com/paperswithcode/releasing-research-code/blob/master/templates/README.md*, 2000.

45 Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, Yingqian Min, Zhichao Feng, Xinyan Fan, Xu Chen, Pengfei Wang, Wendi Ji, Yaliang Li, Xiaoling Wang, and Ji-Rong Wen. Recbole: Towards a unified, comprehensive and efficient framework for recommendation algorithms. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4653–4664, New York, NY, USA, 2021. Association for Computing Machinery.

46 Justin Zobel. When measurement misleads: The limits of batch assessment of retrieval systems. *SIGIR Forum*, 56(1), jan 2023.

4.4 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 [1] 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.

4.4.1 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 [1]:

> 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. 5 and defined here, using the terminology from [2, 1]:

- 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.

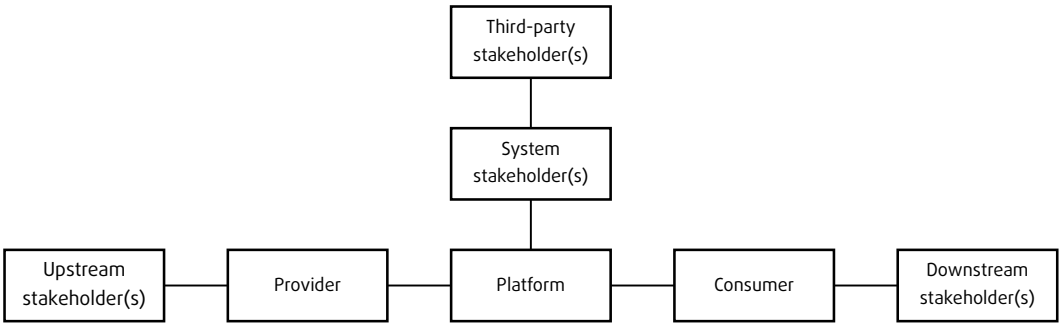

Figure 5 A multistakeholder view of a recommendation ecosystem.

The vast majority of recommender systems research focuses its evaluation only on the perspective of recommendation consumers. However, in most applications, a large number of 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 4.4.4.

Fig. 6 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 [15]. 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 2. 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[13].

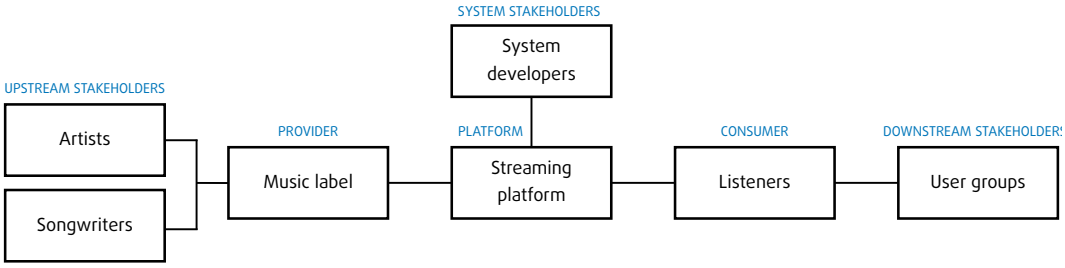

Figure 6 Stakeholder relations for the music streaming example.

Table 2 Sample stakeholder goals and measures for the music 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

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.
- **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.

[13] 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 [57].

- **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.[14]
- **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.

4.4.2 Values

[41] 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 [25]: "*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 [35, 55, 1, 9, 54, 21, 39]. 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 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.

[14] 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.

4.4.2.1 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.

[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, [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." [35] 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.

[41] 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.

4.4.2.2 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 [54, 55].

Usefulness and enjoyment means that recommendations should meet the needs and expectations of its stakeholders effectively and efficiently [28]. 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 [54].

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 [27]. 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 [37]. **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 [22, 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 [52]. See Section 4.2 elsewhere in this report. 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 [44]. 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, 45]. 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 [17]. **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 [50].

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 [58]. 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 [18]. 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 [34]. 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 [35, 41].

4.4.2.3 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 4.4.3, 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 4.4.4, we show how we take theory to practice but discuss several examples of multistakeholder recommender system applications.

4.4.3 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 4.4.4. 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.

4.4.3.1 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. 5. 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 [29], contextual inquiry [46], and co-design [53] could all be beneficial in this process.

4.4.3.2 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 [35]. 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.

4.4.3.3 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 – the reader may refer to Section 4.1 and Section 4.3 for discussions of some best practices and key metrics in recommender systems evaluation. 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.

4.4.3.4 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 [60]. 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, 32] 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 [59]. 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 [19]. 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, 36], multi-criteria decision making [56] (including its various methodologies, such as data envelopment analysis [7], conjoint analysis [22], multi-attribute utility theory [26]), machine learning [40], 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 [30] 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 [43]? 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.

4.4.3.5 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 to incorporate 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. [42] 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.

4.4.4 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.

4.4.4.1 Music Streaming

The first example we consider is streaming music recommendation with the key stakeholders introduced above in Fig. 6, and also included in Table 2.

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 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 receptive[15], 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?

4.4.4.2 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 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 of 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. 7 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'

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

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 4.4.3. 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).

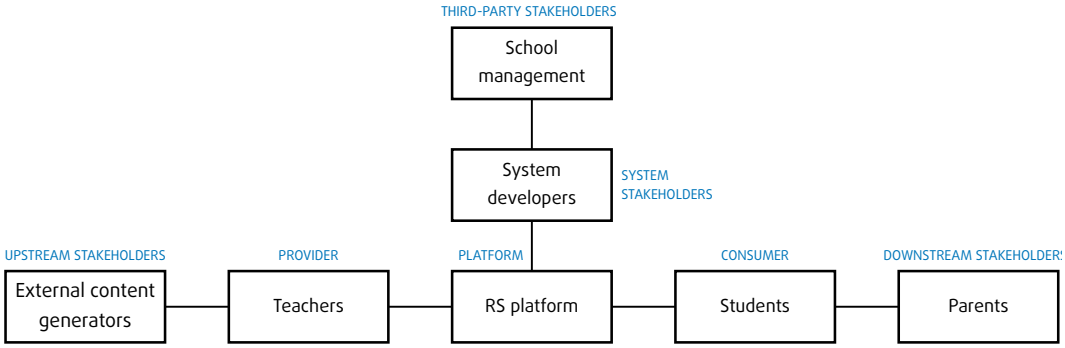

Figure 7 Stakeholder relations for the education example.

Table 3 Sample stakeholder goals and measures 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

4.4.4.3 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 [38] 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. 8 illustrates the different stakeholders involved in this recommendation scenario and is supplemented by Table 4, which displays example goals and measures for each of the stakeholder categories.

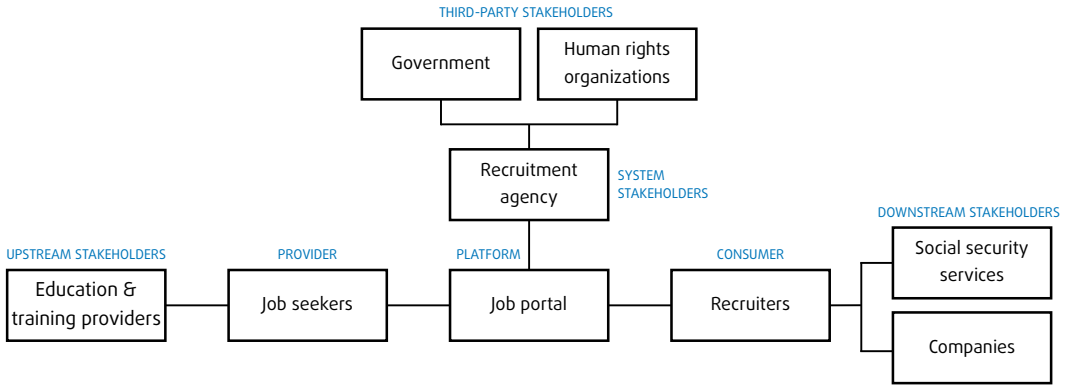

Figure 8 Stakeholder relations for the human resources example.

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

Table 4 Sample stakeholder goals and measures 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

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 metric. 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 [23, 31, 42, 47, 48], although research suggests that these fears can be mitigated by additional AI training [23].

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 [20]. 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. See Section 4.2 for more discussion of recommender systems fairness.

Human rights organizations are non-governmental organizations that seek to 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.

4.4.5 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:

4.4.5.1 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.

4.4.5.2 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.

4.4.5.3 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?[16]

[16] System governance here is different from data governance as discussed elsewhere in this report.

Corporate structures often have a very concrete answer to this question, but as media scholar Nathan Schneider reminds us [49], 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.

4.4.5.4 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[17] 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, 51], we know relatively little about provider-side experiences with recommender system interfaces.

4.4.5.5 User-centric Evaluation

There is nothing in this discussion that requires metrics are behavioral or off-line. [28] 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.

4.4.5.6 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 [24]. 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 [33], but these are solely consumer-side metrics. There is little development of (for example) system-oriented metrics for conversational recommendation.

[17] `https://studio.youtube.com`

4.4.5.7 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 4.4.3.4). There could be the need for new metrics which are explicitly conceived to address the multistakeholder problem and than can be configured to satisfy the different goals selected for the problem at hand.

References

1 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.

2 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.

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.

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.

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.

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, 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.

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 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.

18 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.

19 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.

20 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.

21 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.

22 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.

23 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.

24 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.

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

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

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

28 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.

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

30 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.

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

32 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.

33 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.

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

35 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.

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

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

38 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.

39 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.

40 Aviv Navon, Aviv Shamsian, Ethan Fetaya, and Gal Chechik. Learning the pareto front with hypernetworks. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

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

42 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.

43 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.

44 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.

45 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.

46 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.

47 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.

48 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.

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

50 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.

51 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.

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

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

54 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.

55 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.

56 Evangelos Triantaphyllou. *Multi-Criteria Decision Making Methods: A Comparative Study*. Springer, New York, NY, USA, 2000.

57 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.

58 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.

59 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.

60 Yong Zheng and David (Xuejun) Wang. A survey of recommender systems with multi-objective optimization. *Neurocomputing*, 474:141–153, 2022.

4.5 Evaluating the Long-Term Impact of Recommender Systems

Andrea Barraza-Urbina, Grubhub, USA, abarraza@grubhub.com
Peter Brusilovsky, University of Pittsburgh, USA, peterb@pitt.edu
Wanling Cai, Trinity College Dublin, Ireland, wanling.cai@tcd.ie
Kim Falk, DPG Media, Belgium, kim.falk@dpgmedia.be)
Bart Goethals, University of Antwerp & Froomle, Belgium, bart.goethals@uantwerpen.be
Joseph A. Konstan, University of Minnesota, USA, konstan@umn.edu
Lorenzo Porcaro[18]*, Joint Research Centre, European Commission, Italy, lorenzo.porcaro@ec.europa.eu*
Annelien Smets, Vrije Universiteit Brussel, Belgium, annelien.smets@vub.be
Barry Smyth, University College Dublin, Ireland, barry.smyth@ucd.ie
Marko Tkalčič, University of Primorska, Slovenia, marko.tkalcic@gmail.com
Helma Torkamaan, Delft University of Technology, The Netherlands, h.torkamaan@acm.org
Martijn C. Willemsen, Eindhoven University of Technology & Jheronimus Academy of Data Science, The Netherlands, M.C.Willemsen@tue.nl

4.5.1 Introduction

Recommender systems and recommendation technologies are now a familiar part of the modern information landscape and a routine aspect of our daily lives [23]. Many people engage with recommender systems throughout their typical day – as they plan their morning commute, when they collect podcasts to listen to, when they order lunch, when they pick a movie to relax with in the evening, and if they select a book to wind down with before bed.

There is little doubt that recommender systems are here to stay, and they will continue to play an important role in people's lives as they increasingly influence the media we read, watch and listen to, the food we eat and the exercise we do, the friends we connect with, and even the people we date. In this context, it is important for the recommender systems community to carefully consider the impact of these systems, not only in the short-term (within/between sessions) but also across a time-span that can be measured in months or years. In doing so, it will also be necessary to incorporate richer forms of usage data and external data sources into our evaluation methodologies because click-through rates and rating predictions offer only a limited ability to assess the broader impact of recommendations.

We must strive to understand how these systems will impact all stakeholders in the long term. Doing so will help our research community to have a more positive impact on end-users, provide industry with new opportunities to innovate, and ensure that society as a whole enjoys the benefits of responsible recommendation. Ignoring these issues will likely diminish the value of recommender systems and lead to a skewed understanding of their long-term impact. The latter is especially relevant since recommender systems, like other AI technologies, are increasingly subject to regulatory scrutiny [24].

Consider three common examples of recommender systems and how a long-term perspective can enrich our understanding and design of such systems:

[18] The views expressed are purely those of the author and may not in any circumstances be regarded as stating an official position of the European Commission.

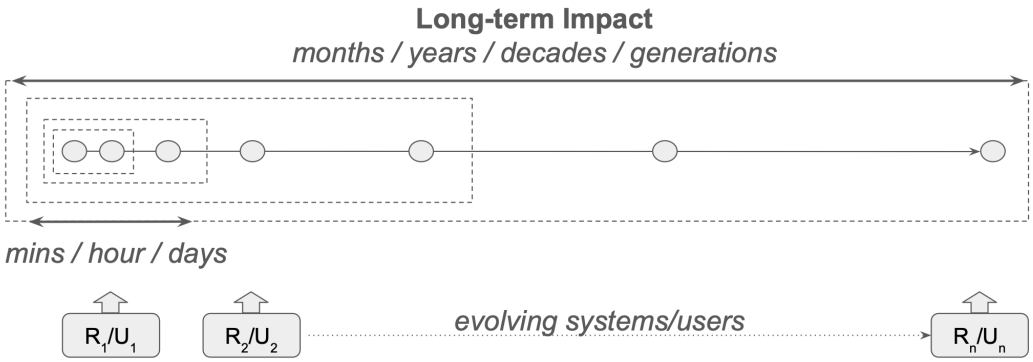

Figure 9 Understanding the long-term impact of recommender systems requires a shift in perspective. While traditional approaches to evaluation can work well to elucidate the impact of recommendations in the short-term, within or across sessions, now that usage can be measured in years or longer there are further opportunities to evaluate the longer-term impact of recommender systems and how this relates to meaningful changes in user behaviour or habits.

1. *An e-commerce site that sends out weekly email recommendations.* A traditional, short-term evaluation of such a system might focus on click-through rates, conversion rates, revenue per click etc. [37], but such an evaluation will be incomplete in several important respects. For example, it will not help us to understand how these emails inform customers about product segments they may not be aware of, or whether these emails lead to future purchases. Moreover, this type of evaluation may not help us to understand whether these emails annoy the customer in a way that limits stickiness, diminishes future visits, or reduces referrals. In other words, short-term evaluations do not shed light on the broader impact that the recommender might have on its users or the e-commerce site. In contrast, adopting a longer-term perspective means that lifetime value, brand reputation, and other factors can enrich our understanding of the impact of such a recommender [88].
2. *A recommender system for encouraging behavior change.* Many such systems have been built to encourage greener choices in energy usage (transportation, home energy consumption, reducing carbon footprint etc.) [75, 76]. In the short term, they can be evaluated based on whether these recommendations are read or bookmarked, or based on how many follow-up actions are taken (e.g., ordering more efficient lighting, requesting further information on home insulation etc.). But the long-term goal must be to change the behaviour and habits of users rather than facilitate short-term transactions. A longer-term evaluation provides unique insights into whether users are making lifestyle choices that are ultimately more sustainable (e.g., reducing their carbon footprint) beyond their interactions with the tool's recommendations. Without that perspective, it will be all but impossible to correctly distinguish between an eager early-adopter whose initial enthusiasm is short-lived and does not translate into more sustainable lifestyle choices, from a more cautious user who comes to recognise the benefits of more sustainable lifestyle choices over an extended period of time. In fact, by some traditional evaluation measures the former may be viewed as more desirable than the latter, and it is only through a longer-term evaluation perspective that the true benefit of these recommendations can be recognised.
3. *A social media recommender designed to keep its users connected, engaged, and informed.* Today, recommender systems go hand-in-glove with social media and the success of many social media platforms has often been attributed to their ability to filter and personalize

content (text, photos, videos etc.) for individual users [52, 29, 20]. Indeed this strategy has been so successful that today a large proportion of people now routinely rely on social media (and their embedded recommenders) as their primary source of news [77]. Such systems are straightforward to evaluate in the short-term: there are numerous examples of studies that have looked at various engagement metrics from click-throughs and ratings (votes, likes, etc.) to read-times (which roughly correspond to advertising revenue) [40]. However, such short-term thinking may lead to systems with serious negative long-term consequences from unhealthy increases in screen-time, to pigeonholing, proliferating hate-speech, and even radicalization. There are concerns that certain groups are particularly at-risk (e.g., teenage boys and girls) when they are bombarded by messages that can have a detrimental impact on their self-esteem and long-term mental health [71]. A longer-term evaluation can actively consider the well-being of users by assessing changes in the diversity of consumption, measures of connectedness to others, and other factors to assess (and design for) the recommender's positive impact on its users.

In the sections that follow, we discuss how a long-term perspective can improve research and practice. We start by looking at how recommender systems evaluation can change to incorporate tracking, collecting, and reporting long-term measures. We then look at the social and behavioral research directions that can support building a better understanding of human behavior, long-term stakeholder goals, and metrics to reflect it. Finally, we look at practice, by examining how short-term thinking can limit or even undermine the potential success of deployed recommender systems, and how long-term evaluation can support the business cases needed to make trade-offs between short- and long-term objectives.

4.5.2 Long-Term Impact and Systems Research

In this subsection, we discuss how long-term impact can be considered in modern system-oriented research in a way that focuses on assessing the performance of recommender algorithms and recommender systems that use these algorithms. We recognize two mainstream types of research – (1) data-driven research, which focuses on algorithm evaluation by engaging available datasets, and (2) user studies, which assess recommender systems by engaging real users. For both types of research, we would like to stress the importance of longer-term studies engaging a broader range of data. In data-driven studies, this may mean collecting and releasing datasets that accumulate user data for several months to several years and include data beyond the limited traditional scope of ratings and clicks. For user studies, it can be achieved by running longitudinal studies and purposefully collecting data that could help in assessing the long-term impact of the systems.

4.5.2.1 Media Recommendations

Consider media recommendation, a very traditional domain for recommender systems, which encompasses many popular and familiar recommendation applications, including music recommendation [38], video content recommendations [54], and news recommendation [45]. The first generation of recommender systems research in these areas focused on available datasets of ratings and assessed the quality of recommender algorithms by measuring their ability to predict these ratings or generate a better ranking list of recommendations [48]. The integration of recommender algorithms in media consumption systems such as Netflix or Spotify, and the ability to collect data beyond simple ratings further extended the range of metrics used to evaluate the systems. The current generation of media recommendation

systems can track how a user responds to a recommendation, for example, by determining whether the user consumed (watched, listened to, or read) a given recommendation, partially or fully, in order to better assess the relevance of the recommendation [40].

Now that many such systems have been in operation for five, ten or even more years, it may be possible to release long-term usage datasets to facilitate evaluation that extends beyond traditional, short-term evaluations and allows for a broader impact assessment (consumption diversity, etc.). Releasing standard consumption/ratings data over multiple years will allow researchers to answer a range of intriguing broader impact questions about how recommendations change or otherwise influence consumption patterns:

When and how much do users consume?
Does consumption variety increase or decrease?
Do users develop new tastes?
Did users discover new types of content that they may otherwise have missed?
Are users becoming more or less satisfied with their media consumption?

In several cases, media consumption systems already collect a broader set of usage data – for example, a movie recommender system can ask whether a viewer is watching alone, with kids, or with significant-other – and tracking this data over time may enable researchers to assess whether the recommender systems help to bring users to spend more time together. Some music recommender systems ask users about their current mood to better personalize recommendations [4, 43]. Releasing these data along with traditional click and rating data will help connect recommendations and watching behavior with long-term mood changes.

In longitudinal studies of recommender systems with target users, the opportunity also exists to collect an even richer range of data by asking users to periodically volunteer various forms of feedback that could be related to a broader impact. This may help better understand how recommender systems affect people's mood, mental health, and general sense of well-being over time. An example of such longer-term studies and the data that these studies could collect is provided by the famous HomeNet project [51], which evaluated the long-term impact of Internet use to identify increased levels of loneliness and a greater sense of isolation among early Internet users.

4.5.2.2 Recommender Systems in Education

Educational recommendation systems (including learning content recommendations and course recommenders) serve as a useful counterpoint to more traditional media recommenders [17]. Research on recommender systems in this domain is still in its earlier data-driven stage, as researchers attempt to assess performance using regular data collected before integrating recommender systems in the application context. This data-driven approach to assess the quality of personalization is typically focused on predicting learner performance when solving a specific problem or during an exam. User studies, which are natural in this domain, can also collect learner feedback on question/content difficulty, novelty, or relevance of suggested items and courses, although it does not necessarily help to assess the longer-term impact of recommendation in this domain.

However, the educational domain benefits from a much broader set of data, which could help assess several dimensions of longer-term impact. Even in relatively restricted online learning content, existing systems collect all data about user interaction with learning content, course discussions and integrated assessments. Using these data, we can assess whether a learner content recommender system has helped to make the learner more efficient (i.e., helped to gain the same level of knowledge faster) or whether it has helped the learner to achieve an improved knowledge level for a given unit of effort? Did the recommendations

help reduce the number of cases where learners needed to ask questions in the forum? Or did they increase the number of cases when they answer questions from peers? Does the use of a recommender system in a prerequisite course help the learner to perform better in a future course that requires this prerequisite knowledge.

In a more traditional context, universities and colleges could collect an even broader set of data covering learners' life beyond courses: exercise, club activity, volunteering, internship, and job placements. These data may help explore the relationship between their approach to learning and their lifestyle: does improving learning efficiency lead to a more satisfied, healthier learner, because they can spend more time exercising and relaxing? We should also be able to assess whether course recommendations helped students diversify their studies, help them become better prepared for the modern workplace, and otherwise improve their employment prospects. In fact, many universities are already collecting this type of data, augmented with various feedback from students (i.e., course feedback, internship reports), and their newly formed "analytics teams" have already gained experience using these data to assess the broader impact of major curricular innovations. This experience could be used to assess the broader impact of educational recommender systems.

4.5.2.3 Combining Multiple Studies

An important aspect of longer-term research is the need to assemble and compare data obtained from multiple studies. A reliable evaluation of longer-term impact in a single study requires a relatively stable set of conditions over an extended period of time and reduces our ability to assess multiple research ideas or options simultaneously. Assembling results from several offline or online long-term studies may enable the research community to more reliably assess the long-term impact of multiple system design aspects. Does the specific recommender approach lead their users to enjoy a more diverse collection of artist and genres? Is it decreased or increased their overall listening? Are they more or less satisfied with how much time they spend for music listening or movie watching? Does a novel transparent interface with better user control made the user return to the system more frequently or, in contrast, pushed the users to use other systems? Which combination of algorithms and interfaces in a course recommender get their users better prepared and more satisfied with recommendations in the longer term?

In turn, the need to compare and integrate data from multiple studies makes it more important to agree on the set of long-term focused data to be collected and a set of long-term impact factors to measure. Moreover, in order to enable this type of meta-analysis, the recommender systems community will need to evolve its approach to evaluation to adopt a level of experimental rigour and reporting standards that facilitate such opportunities; see Section 4.5.5.

4.5.2.4 Developing Long-term Data and Metric Sets

Part of the challenge of conducting long-term research (whether retrospective analysis of data sets or experimental user studies) is that new ideas and phenomena arise for which the collected data or experiment design are inadequate. For example, a dataset collected to study the quality of recommendations may not have captured data that would allow assessing the diversity or unexpectedness of those recommendations. An experiment looking at the effects of different recommender algorithms or interfaces on consumption may not have baseline data on user attitude towards the recommender or brand. Accordingly, there is an increasing need to develop standard suites of metrics and data sets to support such long-term impact research.

It is beyond the scope of this section to specify the contents of such a standard – rather we make the case that researchers in the field should promulgate and evolve such standards with a goal of converging to a relatively comprehensive set (and note that several other sections of this article, including the section immediately following, propose partial solutions). We suggest that some factors to consider include:

1. The challenge of identifiability of users in the context of such data, and therefore the possible need for explicit informed consent. (Consider for example [7].)
2. The desirability of preserving not only user interactions, but also the system prompts that lead to those interactions (e.g., recording displayed recommendation sets).
3. Recording a set of interval metrics on a regular schedule (e.g., periodic logs of consumption properties, recommendation properties, logins, etc. for the past week). (Consider [49] or [14].)
4. Developing a suite of general attitudinal and beyond-system behavioral survey questions that can be administered regularly (subject to the appropriateness within the system context).
5. *If you want to measure change, do not change the measure.* In order to avoid issues with inconsistency over time, statistical validity, reliability across waves, bias introduction, and introduction of confounds, the metrics used should be well-thought upfront and not changed throughout the long duration of the study.

4.5.3 Social Behavioral Research with Long-term Impact

The collection of long-term data and their respective metrics, as discussed in the previous section, can benefit from being informed by insights from social behavioral research. Sometimes we just want to be entertained (or distracted); other times, we want to develop a new taste in music, improve our fitness through exercise, or pursue other *long-term goals*. Our behavior is driven by both short- and long-term goals, but we often procrastinate and prioritize immediate gratification, lacking the self-control to achieve our long-term aspirations, goals and preferences. Many recommender systems are predominantly focused on fulfilling such short-term immediate needs and desires, being optimized and evaluated only in the short term. Furthermore, many recommender system goals are not only short-term but also business-centric rather than user-centric.

In order to address long- vs. short- and user- vs. business-centric evaluation we look from the perspective of the user and society. In particular, we discuss social science-informed theories that explain user and societal behaviour to operationalize long-term impact metrics. Next, we discuss how social sciences understand how long-term goals can be achieved and how this informs the evaluation of long-term impact from the perspective of achieving long-term goals.

4.5.3.1 Metrics from social sciences to understand and evaluate long-term interactions

Systems are learning about preferences and behavior while interacting with users over time and thus have long-term impact. Companies optimize their recommender systems for business metrics, but they do not necessarily account for other impacts (e.g. Netflix optimizes for hours of viewing but is not aware if these hours are quality time or addictive binge-watching [18]). Binge-watching, for example, has been studied in psychology and has been related to mood regulation [73]. Hence, understanding how mood regulation works can inform the choice of metrics to be used for measuring the impact of a recommender system.

The choice of metrics is domain-dependent. Here we provide a non-exhaustive set of theoretical concepts relevant to the impact of recommender systems on both individual users and society. The concrete metrics used need to be adjusted to the specific domain. These theories primarily draw from psychology and other social sciences. Furthermore, there are additional relevant theories from economic, cultural, behavioral, and various other fields that can also be considered to fully understand and measure the long-term impact of recommender systems.

Individual user behavior can be better understood if we recognize that users differ substantially in their personal characteristics. Some of these characteristics are hard to change (e.g., personality) while some others can be affected by the exposure to recommender systems. For example, the user's level of expertise in a domain, their personality [78], their decision-making style [6], and their need for autonomy/independence. As an example for expertise, the Music Sophistication Index (MSI) [59] has shown to be a substantial factor in understanding individual differences in user interactions with the music recommender [27]. On a more specific level, their momentary behavior will be affected by their attitudes, values, and beliefs: these are relatively stable in the short term but might drift over the course of time, potentially influenced by the interaction with a recommender system. There are several models describing how these aspects influence current behavior, such as the Theory of Planned Behavior [2]. In light of our perspective on long-term evaluation, this suggests that measuring attitudes, beliefs and values on a regular basis might help us better understand what is driving users' long-term interactions with a recommender system. Similarly, users' mood [79] or mental well-being might fluctuate over time and play an important role in their interactions, and any measures that might capture these implicit or explicitly [84, 81, 83] would be helpful in better understanding long-term interactions.

A special psychological construct relevant for recommender systems research is the concept of user preferences. Recommender systems take user preferences as somewhat stable and measurable, but psychological research has shown that people often do not really know what they like and construct their preferences while making decisions [11, 34, 25]. A recommender system thus also allows people to better understand their preferences, and as the recommender system learns over time, the users might also learn about their preferences from the interaction with the system. Moreover, research makes the distinction between actual (current) preferences versus more ideal (value-based) preferences [54, 46], which is directly related to the distinction between short-term and long-term goals, which we will discuss in the next section.

Social theories are crucial in understanding the broader implications of recommender systems on collective behavior and societal structures. These theories can, for example, help understanding how public sentiment, political polarization, and education and awareness might be impacted by recommender systems. Social theories and their analytical frameworks can also help study the effects on societal tolerance, diversity, and potentially economic inequalities or environmental sustainability. Concepts, such as cultural identity, social capital, and civic engagement can also be examined, providing insights into how recommender systems can shape or impact social norms.

Some impact metrics can be computed from logs (e.g., hours of reading news), while others require more specialized instruments (e.g., measuring user sentiment toward a political issue in news recommender systems). These instruments, such as lengthy questionnaires, can be costly and time-consuming to administer. Therefore, a balance between accuracy and scalability is essential. One option is to measure just a small sample of users and build a predictive model for the remainder of the population under study. For example, asking a couple of hundreds users to gather ground truth labels and then training a predictive model from user behaviour logs.

To measure how a news recommender system affects political polarization, one could sample the current sentiment of the society on a regular basis (e.g., weekly) over a longer period of time. The theories that inform the choice of the metric could be, for example, the social identity theory [85] (in-group favoritism and out-group hostility) and the cognitive dissonance theory [28] (reject or rationalize information that contradicts people's beliefs). These theories could lead to a choice of metrics, such as measuring the sentiment of people towards in-group and out-group generated posts (can be computed from digital traces in social media) or a questionnaire-based instrument that measures the cognitive dissonance of users when/after being exposed to a certain news item.

4.5.3.2 Supporting short- and long-term goals of users

Psychology has studied extensively the conflict between short-term needs and desires and long-term aspirations and goals and how to help people overcome their short-term desires to focus on the long term. Models of behavioral change talk about different stages in which users go from awareness to motivation to change to action (e.g., the transtheoretical model [69]). Are we able to capture such stages in the data and develop metrics for them?

An effective approach to achieve long-term goals is to break the long-term goal into smaller and attainable short-term goals. Such short-term goals have a prospect towards attaining the larger long-term goal, but most recommender systems do not have a notion of a long-term goal being behind the interaction / behavior of the user. Some exceptions are Rasch-based recommender systems [75, 72] and other approaches [82, 10], which models the user's ability and item difficulty, allowing the system to recommend items that are within their ability, thus allowing for smaller short-term goals (I can run 5km now) to be achievable and to develop towards attaining long-term goals (I want to run a full marathon). In any case, recommender systems might need to be aware of such long-term goals, and it is quite likely that we cannot learn about such goals by just observing user behavior with the system. A conversational approach between the system and the user might be needed to make sure that what the recommender system is learning about the user reflects the user's longer-term goals. But how should the system communicate to the user what it learned and based on what metrics? There is an opportunity to develop algorithms that take into account the balance between optimizing for short-term (attainable) goals while still being on track to achieve users' long term goals. What metrics would we need to optimize for both?

There is an inherent temporal aspect in distinction between long- and short-term goals. Long-term goals are by definition more into the future, though they might influence current (short-term) decisions. Research on inter-temporal choice shows that we devalue future gains and prefer immediate rewards over delayed ones and that users need to have awareness of their future goals to overcome this, for example by changing their perspective. For example, we can prevent people from procrastinating by making the goals explicit, making people think more about their future selves, or reserving their mental queries [44]. How can we build recommender algorithms that support such strategies, that can recognize items that satisfy long- and short-term goals and based on what metrics?

4.5.3.3 Cross-fertilization between social science research and recommender systems research

Better measurement and modeling of long-term interactions between users and recommender systems also offer opportunities for cross-fertilization between social science disciplines and recommender systems research. For example, social science research in behavioral change has

shown to have limited practical impact because often studies are designed for understanding and theorizing rather than really helping users move forward. Actual intervention studies are typically done on a much smaller scale compared with large-scale recommender systems experiments. Moreover, these studies typically do not use highly personalized interventions, as they lack the (long-term) behavioral data and algorithm expertise to do so. The outcomes from better long-term recommender systems evaluation studies can inform future intervention studies. Furthermore, recommender systems researchers could team up with domain experts to build multidisciplinary teams to combine the strengths of both worlds.

4.5.4 Long-term Impact in Practice

When designed and implemented thoughtfully, recommender systems can create significant value for their users, providers, and other stakeholders. In this section, we discuss the long-term impact of recommender systems in practice. We first outline the most common (short-term) metrics used in the evaluation of recommender systems, and the potential pitfalls of using these metrics regarding the long-term effects. Going from there, we discuss what more ambitious recommender systems could entail and why, in practice, this may (still) include short-term metrics to measure the performance towards long-term goals. Finally, we conclude this section by discussing some examples of recommender systems research wherein long-term and longitudinal aspects have been considered.

4.5.4.1 Current Challenges: Unintended Impacts of Short-term Metrics

Typically, recommender systems are used in domains where there is an abundance of products or items, helping users discover new content that they might not have found otherwise. This way, not only do the most popular items get visibility, but the long tail of less popular items and niche content can also find its way to its specific audience.

Ideally, recommender systems are beneficial for the long-term goals of its providers, improving overall engagement, retention, and increased revenue for commercial providers [23]. In practice, recommender systems are often optimized using several short-term metrics focusing on immediate user interactions and engagement. Unfortunately, such short-term focus does not necessarily correspond with long-term benefits or might even become harmful for it. Here are some example metrics commonly used.

- Clickthrough Rate (CTR): One of the most commonly used metrics that measures the percentage of recommended items that are clicked by users. It is a direct indicator of how engaging or relevant the recommendations are perceived to be. Systems optimized for CTR, however, can suffer from clickbait. These are recommended items that typically have sensational or misleading titles or images designed to attract clicks. While such content might generate high immediate engagement, it is typically low in quality and does not provide lasting value to users. This can significantly degrade the overall and long-term user experience on the platform. Moreover, such clickbait is likely to result in a feedback loop where the recommender system will put even more emphasis on such sensational content, limiting the diversity of content that users are exposed to.
- Conversion Rate and Monetary Value: This metric tracks the percentage of recommendations that lead to a desired action, such as making a purchase, signing up for a service, or the economic value that it brings. Conversion rates are typically used for e-commerce and service-oriented platforms. Recommender systems optimized for conversion or monetary value might put too much focus on high-price or high-margin items, which might not be

the most optimal choice for long-term monetary value. For one, this metric does not take the post-purchase experience into account and users could lose trust in the system in the long term.

- Immediate User Feedback: This can include thumbs up/down, star ratings, shares, comments, or any other form of quick feedback that users provide after interacting with recommended items.
- Session Duration: This measures the total time a user spends on the platform during a single session. Longer session durations typically indicate that the recommendations are engaging users effectively. Although session duration is very closely related to CTR, it could improve the quality of the recommended content, as CTR alone does not capture the time a user eventually spends on the clicked item.
- Bounce Rate: The percentage of users who leave the platform after viewing a single recommended item. A lower bounce rate suggests that users are finding value in the recommendations and choosing to explore more content.
- Item Coverage: This measures the proportion of the catalog that is recommended over a period of time. Higher item coverage indicates that the system is leveraging a broader range of available content, which can be beneficial for both users and content providers. This is one of the main strengths of recommender systems, activating the long-tail of the catalog and matching niche items to the users interested in it [41].
- Hit Rate: The proportion of times the recommended item is the one that the user interacts with. This is a straightforward measure of the accuracy of the recommendations.

These metrics are essential for understanding how well a recommender system performs in the short term and are often used to guide iterative improvements and A/B testing. However, while these metrics are useful for immediate optimization, they all share the risk of reduced content diversity, and pressure on content creators and vendors to optimize their offering solely to boost the used metric.

Apart from content diversity, there are significant risks of several other long term effects that need to be considered when deploying a recommender system.

- Filter Bubbles, Echo Chambers, and Polarization: Recommender systems can create filter bubbles, where users are only exposed to content that reinforces their existing beliefs. This can lead to echo chambers, reducing exposure to diverse perspectives and potentially fostering polarization [57, 5, 65].
- Addiction and Overuse: Systems optimized for short-term engagement can encourage excessive use, leading to addiction. This is particularly concerning on social media and video streaming platforms, where the continuous feed of recommended content can lead to unhealthy consumption patterns.
- Bias Amplification: Recommender systems can amplify existing biases present in the data. For example, they may disproportionately recommend content from certain demographic groups or types of content, reinforcing societal biases and inequalities.
- Privacy Concerns: Long-term data collection for improving recommendations can raise significant privacy concerns. Users may become uncomfortable with the amount of data being collected about their preferences and behaviors over time.
- Content creators: Recommender systems can skew visibility and revenue opportunities towards already popular content creators, making it harder for new or niche creators to gain traction (see also Section 4.4. This can lead to a lack of diversity in the available content and reduce the overall variety and innovation within the content ecosystem. This situation limits the exposure of different types of content and discourages new creators from participating, which negatively affects the richness and dynamism of the platform.

- User Manipulation: By optimizing for engagement or sales, recommender systems might manipulate users into behaviors that are not in their best interest, such as overspending or engaging with misleading information, or even causing emotional impacts [80].
- Reduced Serendipity: Over time, users may be less likely to encounter unexpected or novel content that could enrich their experience, leading to a more monotonous and less stimulating interaction with the platform.

4.5.4.2 Can we do better? More ambitious recommender systems

Given the problems stated in the previous section, we suggest taking a more ambitious approach to evaluating recommender systems in practice. In this section, we discuss what more ambitious recommender systems could entail and in the next section, we go into more detail of how to proceed with implementing such goals in practice.

In this pursuit, those who are responsible for the roadmap of these systems should recognize the relevance of these long-term objectives. Within the domain of science and technology studies, the social construction of technology (SCOT) emphasizes the notion that technological systems, such as recommender systems, are influenced not solely by technical elements but also by social processes and human decision-making [47]. In the context of recommender systems, this perspective highlights that these algorithms are designed and implemented based on specific choices made by developers, product managers, and other stakeholders [74].

While public discourse often portrays (the impact of) recommender systems as inherently "bad" or problematic, for example, the highly popularized filter bubble hypothesis by [65], this SCOT lens reminds us that the perceived issues or biases in these systems stem from the underlying human decisions and values embedded in their design and development. In other words, recommender systems do not operate in a vacuum but reflect the priorities, assumptions, and trade-offs made by the individuals and organizations responsible for their design and implementation.

However, this does not imply that there could not be any (positive or negative) unintended consequences. For example, as pointed out by [87]: "a recommender system designed to serve its customers may unintendedly (and systematically) contribute to filter bubbles and echo chambers [...] although that was never intended by its designers."

Apart from these unintended consequences, a goal-oriented recommender design in practice is informed by answers to the questions:

"What kind of recommender systems do we want to develop?"

"What kind of objectives do we want to achieve?"

This is where it gets challenging. To answer these questions, one should have a thorough understanding of both the specific domain and the various purposes that may or may not be achieved by using recommender systems.

In most cases, stakeholders either are domain experts or recommender systems experts. These domain experts have the knowledge to define the long-term goals, which should then inform the evaluation criteria of the recommender system. In practice, this translation from "goals" to "metrics" [35] is a joint effort by domain experts and recommender systems engineers, which is mediated by product owners who facilitate the interaction between these stakeholder groups. In this effort, it is essential that each of these stakeholders is informed about the range of possible long-term goals of these systems. This is to ensure that they are not constrained by the narrow range of objectives that have been dominant thus far in recommender applications in practice, as previously outlined in this section.

To build a broader, or more ambitious, understanding of such long-term recommender goals, practitioners may rely on examples discussed earlier in this report or recent surveys [37, 36]. Additionally, a north star in this context could be the UNESCO four core values (Ethics of Artificial intelligence) [86] (See also the discussion of values and goals in Section 4.4.2):

- Human rights and human dignity – Respect, protection, and promotion of human rights and fundamental freedoms and human dignity
- Living in peaceful – just, and interconnected societies
- Ensuring diversity and inclusiveness
- Environment and ecosystem flourishing

While these should be the guiding stars for all AI technologies, optimizing for or even understanding how one system can impact these values might not be easy. Instead, we propose to focus on more specific goals in their respective domain. In the following section, we will look at more specific cases.

A currently open question is: What are the effects of the explosion in social media usage and the new online lifestyle? However, most research [32] seems to indicate that the current level of addiction should be limited, especially for young people but also for the population in general. It is, therefore, prudent that recommender researchers and engineers start thinking about the long-term impact of their systems.

Monitoring how the users are progressing towards the long-term goal could be done simply by asking the user. For example, Duolingo has done extensive work to understand how to measure long-term effects with short-term metrics [30] and adds quizzes and review exercises to understand the progress of its learners better [68].

Another approach could be to request domain experts to define measurable proxy metrics in a shorter time frame, enabling the engineers to optimize the system accordingly. For example, focusing on customer lifetime value could simply be reduced to optimizing users' chances of returning to the platform [30].

There is often a discrepancy between ideal and actual preferences among users and providers of recommender systems. For instance, while the nutritional benefits of broccoli are well known, recommending it may diminish trust in the system because users may prefer crisps instead. To ensure long-term effectiveness, recommender systems should prioritize optimizing for long-term goals. In the aforementioned example, instead of consistently suggesting crisps, an optimization strategy could be to gradually increase the instances where the user selects the ideal choice without causing them to abandon the platform.

Similarly, one might also consider the addiction-like problems that users experience with social media platforms, where the platforms' optimization criterion is to keep users engaged as frequently and extensively as possible [32]. Instead, these social media platforms could set goals to encourage and support physical meetings and events. Could it even be advantageous for these platforms to have users spend shorter but more focused time on their platforms? This approach could provide a more effective platform for marketing, as it might engage focused users rather than relying on the large percentage of mistaken clicks that currently inflate the platforms' metrics. The success of such an approach could be measured with check-in-like features that allow users to demonstrate that they met in person.

Social media has also become many people's main news source, giving a unique chance to provide complete news coverage not only covering many diverse stories but also with opposing views of stories (from different newspapers), ensuring that a user gets exposed to a wide set of topics. While it might not be possible to provide opposing views of stories for a single newspaper, they could still adopt similar goals. Editors should define their overall coverage goals, and a short-term metric would be to optimize so that users would get as complete coverage as possible.

4.5.4.3 Proxy short-term metrics for long-term goals

One of the biggest challenges for businesses today is to define the tests that will allow them to understand long-term impacts better. In practice, the first task would be to define these long-term goals and align those with the stakeholders. The long-term goals should then be translated into Overall Evaluation Criterions (OEC) to ensure they are measurable [50]. As these long-term goals often require measurements across a longer period of time, one might be tempted to suggest that online tests should simply run for longer. However, often, that obstructs other goals of stakeholders, and even if allowed, this does not come without its pitfall either [16].

Another more feasible approach is to create a set of short-term proxy metrics that will enable performance measurement towards long-term goals but in short-term feedback loops. Using proxy metrics is not always straightforward, as described by [63]. It is, therefore, important to capture the proxy metrics and compare them with the long-term goals at intervals and monitor the overall system to ensure that using these metrics won't hurt the system.

Research shows that diversifying recommendations increases the user experience. This is not the best strategy when optimizing for short-term goals, but by diversification, the system learns new user preferences [82], and even if it might result in lower short-term performance, it will be a good investment for the system's long-term performance.

Lastly, most metrics reported only look at the positive increase, but this can very well hurt minorities, as they might be hurt badly by changes, which might still go into production because an A/B test is considered successful. Similarly, another metric seldom considered is the churn rate of users who receive recommendations that affect them to the point that they leave and never return. Most short-term metrics optimize for positive reactions, while bad recommendations are never tracked and monitored. Not doing this will eventually lead to a loss of users. Returning to the examples of broccoli vs. crisps, it is important to show recommendations for broccoli to encourage healthy behavior, but not to the level that makes users not return to the platform.

4.5.4.4 Examples of long-term impact research

This section presents few examples from the recommender systems literature wherein long-term and longitudinal aspects have been considered when assessing the impact of the recommendations. First, there are presented works employing two different methodological approaches, simulation-based environments and longitudinal user studies. Then, examples of studies which consider specific scenarios emerging from continuous interactions with recommender systems are described: feedback loops, impact of content diversity, and rabbit holes.

Simulation based-environments, such as Agent-based Modelling (ABM) has been employed to explore the long-term impact of recommender systems on various aspects [1]. [88] focus their efforts on simulating users' consumption strategies, demonstrating how, through reliance on recommendations, individuals might inadvertently contribute to a decrease in overall variety over the long term. [89] utilize ABM to investigate the impact of preference bias – the distortion in users' self-reported ratings resulting from recommendations – on the effectiveness of recommender systems. Specifically, they demonstrate how the system's performance can be adversely affected by the introduction of user-rating-induced bias, potentially compromising the overall variety of recommended items. [39] concentrate on the analysis of recommendation techniques through an iterative approach, where users are

presumed to engage with a particular portion of the recommended items. They demonstrate that, in terms of recommendation dispersion and coverage, several systems evaluated exhibit an increased concentration over time. Employing a similar methodology but focusing on session-based recommender systems, [26] uncover similar findings concerning spread and coverage.

Longitudinal user studies are quite rare in recommender system research, due to the large amount of resources and time needed in order to gather data on the interactions between users and systems. An eight-week longitudinal study between subjects has been conducted by [15], designing an app where participants received personalized recommendations for physical activities and guidance to minimize sedentary behavior. In the work by [33], the impact of personalized recommender systems is examined. The system provides visual feedback and recommendations based on individual dietary behavior, phenotype, and preferences. By employing quantitative and qualitative measures over a 2-3 month period, the study demonstrates that the system positively impacts nutritional behavior as measured by the optimal intake of each nutrient. In the music field, [53] present a longitudinal study, focusing on users' exploration behavior and change in behavior after employing a music genre exploration tool for four sessions across six weeks. [67] present the outcomes of a 12-week longitudinal user study, involving participants who received daily music diversified recommendations. By analyzing their explicit and implicit feedback, it is demonstrated that exposure to particular levels of music recommendation diversity in the long-term may impact listeners' attitudes.

The decisions made by recommender systems can shape user beliefs and preferences, which subsequently impact the feedback the system receives, thereby establishing a long-term **feedback loop**. [42] provide a theoretical analysis of the relationship between feedback loops, echo chambers, and filter bubbles. [13], through the simulation of various user engagement models with recommender systems, demonstrate the influence of feedback loops on the homogenization of users' behaviors. [56] design a model to iteratively analyze the feedback loop, showing how it may be responsible for a decline in aggregate diversity. Focusing on the long-term impact on exposure, [21] discuss how recommenders may exacerbate the rich-get-richer effect, strengthening exposure inequalities. Challenges derived from the presence of feedback loop are also common in industry settings, and [85] show how to address long-term feedback loop emerging issues by using an offline evaluation framework.

Content diversity has been at the center of attention of numerous studies due to its relationships with filter bubbles and echo chambers, among the undesired long-term impacts most researched in the recommender system literature [57]. [8] employ numerical simulations to model user decision-making processes, offering an explanation for the findings of a prior study by [60] on the impact of recommender systems on content diversity. In the original work, the authors found that users engaging with the provided recommendations consumed more diverse content compared to those who did not. [8] corroborate these results, but they also report an increase in user homogeneity – a decrease in aggregate diversity. Similar results are presented also by [3], who observe a connection between recommendations and long-term reduction of diversity. The narrowing of the range of content to which users are exposed is also relevant when the pathways that recommender systems define lead to the consumption of polarized content – eventually contributing to user radicalization – creating what are nowadays commonly referred to as **rabbit holes** [64]. Under this lens, YouTube recommendations are examined in the work by [70] and [22] in the context of user radicalization.

4.5.5 Towards More Rigorous Experimental and Empirical Research in Recommender Systems

4.5.5.1 Introduction and Motivation

The goal of experimental and empirical research is to contribute new knowledge that future researchers and practitioners can use and build on with confidence. Fields of research generally rely on two mechanisms for ensuring that proposed contributions deserve that confidence:

- Peer review – the evaluation of work by other experts in the field
- Standards – the adoption of practices viewed as best practices for research

Consider, for example, a medical researcher who wants to test whether high doses of vitamin C affect the incidence of influenza among people who take it. Standards exist for clinical trials to constrain the methods (e.g., protocol development, funding approval, protocol registration, a double-anonymous, placebo-controlled, random assignment trial) and the parameters of the experiment (e.g., through power analysis and pre-determined significant effect sizes) [58]. A peer review process would likely be applied twice – once beforehand of the study design (either as part of a funding decision or as part of human subjects ethics review), and then again afterward on the final manuscript. Even then, standards of publication would ideally ensure that sufficient detail be included in the publication to support both replication and later meta-analysis for assessing the impact of previous research studies.

In recommender systems, like many other computer science-related fields, in contrast to the medical domain mentioned above, our mechanisms for ensuring confidence in results have been more limited. We have peer review prior to publication, and have numerous best practice guidelines published (e.g., [14, 48, 49, 38]). We also have ACM guidelines for conducting studies with human subjects and the need for institutional IRB. But there are no accepted or agreed on standards for reporting results, no pre-review of experiments (as a way to demonstrate that the hypothesis, approach and analysis were planned in advance and not shaped by data as they emerged), and rarely any mandate to authors or reviewers to reference best-practice guidelines as part of publications and their review. We should note that recommender systems research benefited significantly from this flexibility in its early years. Exploratory work like the early recommender systems implementations needs rapid exposure more than iterative refinement. Indeed, there still is and always will be highly exploratory new work. But we believe that the majority of research in the field is incrementally advancing the science and practice of recommender systems and would therefore benefit from increased focus on rigor.

In this section, we make recommendations aimed at improving research rigor and the confidence with which research results can be applied. While some of these recommendations are general and can apply to any empirical or experimental research, we focus primarily on high-cost research (such as longer-term and large-scale experiments, but also computationally expensive multi-dataset experiments and simulations) where peer review of research design may help address study design problems in a timely and cost-effective manner.

4.5.5.2 Case study of our proposed approach: Special Track for Registered Reports

Overview and structure. This track would serve as an implementation of registered reports for ACM Transactions on Recommender Systems (TORS) and would serve as an example to evaluate for possible future implementation in other venues including the ACM RecSys conference. The concepts of registered reports and preregistration have been popularized in

health and social science fields (in part in response to concerns about the replicability of prior published research); see [31, 9, 61, 62, 55, 66]. The key element of registered reports is the separation of research and peer review into multiple phases. The researcher designs a study and writes up that design (the research protocol or plan), then peer-reviewers review that design (perhaps requiring changes). Only after the design has been accepted does the researcher carry out the study followed by a more streamlined peer review of the resulting publication. The goals and benefits of this mechanism are twofold:

- Reduce the risk of experimenters changing their designs as an experiment proceeds to steer towards positive results (e.g., "*gee, it doesn't seem like click-through is improving, let's look at some other metrics*").
- Ensure the design will inspire confidence in the results by using peer feedback to modify it before the study (which is cheap) rather than delaying peer feedback until afterwards (when it is expensive or impossible) (e.g., "*gee, this would have been a really good study if you'd pre-tested all your users before they experienced the recommender – can you go back and do that?*")

In this section, we outline the intended scope for this track, instructions for researchers and reviewers, and notes on how the track would operate. TORS already has publicized its willingness to publish registered reports, and this proposal has been developed in consultation with the editors. We should note that we advise recruiting a carefully-selected set of proven reviewers and editors to launch this track as its success depends heavily on the quality and timeliness of reviews and the researcher experience.

Intended Scope for this Track. These are research studies that submit proposals (detailed experimental justification and design) prior to conducting the study. **The submission for this track is not intended to be the primary form of publication and review. Instead, this serves as a platform for high-effort and high-cost work with the anticipation of high-value outcomes.** In some ways, this approach echoes the dissertation proposal model of PhD programs, aiming to inspire confidence in PhD students to tackle significant questions over an extended period (typically 4-6 years) with proper feedback and assessment. It is also closely related to registered reports and some practices from the medical field (clinical trial reports).

What type of submission may be expected for this track Example:

- Large-scale user experiment to evaluate the impact of different recommender techniques
- Longitude studies to investigate the long-term impact of recommender system design
- Expensive and/or time-consuming dataset-based studies

What type of submission may not belong to this track Example:

- Experimental studies that have already conducted experiment and obtained the results
- Studies that perform typical offline evaluations of new proposed recommendation techniques on one or more datasets
- Studies that propose new evaluation methods or new metrics as their primary contribution

Submission Guideline for Researchers. In this Registered Reports track, we encourage researchers to submit proposals for experimental protocols, including detailed experimental justification and design, prior to conducting the study. This aims to collect peer feedback to help researchers ensure the appropriate design before conducting the high-cost study. To enable reviewers to provide constructive feedback on the experimental design, researchers need to cover the following points in the submission:

- Clearly indicate the motivation and objectives of conducting the planned research grounded in the understanding of prior work in the field

- Present specific research questions and/or hypotheses (including the primary hypothesis and secondary hypotheses, etc). in the planned research
- Provide a thoughtful review of the background/context of the research, including the applications, techniques that would be used, and/or the stakeholders that are expected to be involved.
- Present detailed methodology for the planned research
  - Present methods/theories/techniques/applications that are appropriate to the research questions, hypotheses, and questions, and justification for the used methods.
  - Describe the selection of appropriate measurement/metrics for the experiment, including behavioural measurements and psychometric measurements.
  - Clearly state the proposed study design of the research, including consideration of study design factors such as randomization, assignment, experimental tasks, data collection methods, etc; explain why the particular study design has been chosen in preference to other possible designs (i.e., justification for the choice of study design).
  - Describe the procedure for conducting the planned experiment, including
    - Recruitment of participants
    - Randomization, assignment, and bias mitigation (e.g., how to manage dropout and resulting bias which may be present in longitudinal user studies).
    - Study procedure, e.g., what will happen to participants once they are enrolled in your study, how to collect and process data. (Please note: if there are any points in your study where you plan to check interim results and possibly make changes, these must be planned explicitly in this proposal. Changes that are not anticipated in the proposal will result in the study not being acceptable for publication.)
    - Data collection, e.g., how the data will be collected to answer the research questions and verify the hypothesis of the study (e.g. questionnaire, behavior data logged in the systems)
  - Describe the statistical considerations and data analysis methods.
    - Having a prior estimate of effect sizes, power analysis, and what would constitute meaningful and significant results;
    - Stating a specific statistic or method that would be used for analyzing the data being collected. This may be dependent on the data collected, but we do not want fishing around for results.
- Expected outcomes from the research and a brief plan for results report
  - Expected outcome may describe what deliverables (e.g., artifacts, impact of studies techniques on individual users) would be provided in the results
  - In terms of results reporting, a brief plan may include how to follow best practices for how the results will be calculated and reported (e.g., how do you handle cases where the algorithm makes no recommendation or prediction; how do you compute population means across individual metrics)
  - Be specific about datasets and statistics to be released. The goal should be to provide sufficient information to support both replication studies and the use of your results without the need for replication.
- A timeline for the planned research experiment, detailing the schedule from design to completion and publication, or termination under reasonable conditions and within an appropriate time span.
- Consideration of ethics and best practices for responsible treatment of participants and stakeholders.

- For example: it is not often appropriate in many domains to assume long-term goals for participants rather than allowing participants to articulate their own goals.
- Benefits for participants in participating in the studies (if any)
- Safety concerns: provide adequate information on how the safety of research participants will be ensured (if the research may induce some risks to participants).

What happens next? Your proposal will be reviewed by a set of experts. Possible results of the review are:

Accept. This paper will be accepted for publication if you carry out the research in accordance with the proposal. Please remember that you cannot change the research design along the way – any changes would either need to be submitted as new proposals (with the experiment re-starting) or would result in a paper that you would need to submit to a different track or venue.

Accept with Conditions. This paper will be accepted for publication if the results meet the conditions provided by reviewers. Typically this type of acceptance is used when a proposed study would only make a significant contribution to the field if certain results are found, but not if they are not found.

Revise and Resubmit. The reviewers feel your proposal has merit, but require changes to it. Please address those changes (typically method changes or preliminary work) and resubmit.

Reject. The reviewers do not feel your proposal is suitable for this track. This could be the nature of your work, the expected results, or other reasons. We encourage you to consider whether this work should be pursued, and if so to submit it to a different venue or track.

If your proposal is accepted (with or without conditions), it will be published as part of the TORS registered reports registry. [Note: It is an implementation detail to be determined as to whether that registry is part of the ACM DL or uses an external site such as COS.]

Instructions to the reviewers. Thank you for agreeing to review in the Registered Reports track for ACM TORS. Registered reports are research studies where the study design is evaluated prior to conducting the research.

Guidelines for Reviewing Registered Report Submissions

The intent of this model is to improve the research studies (by improving the design while it can still be changed) and in turn produce more rigorous, reliable studies. Another goal is to improve the experience for authors who can get timely feedback and avoid wasting time on work that would not meet publication standards.

As a reviewer, you will be asked to perform a pre-review of research designs. For those designs that receive favorable pre-review, you may also be asked later to review the final paper submissions (post-review) to verify that the work adhered to the design and me the criteria for the review.

We want to note up front that registered reports are not necessarily appropriate for all types of research. They are generally most appropriate for high-effort research (such as experimental studies) that are designed to produce reusable research results. Exploratory studies, quick studies, case studies, and other forms of research usually don't fall into this category. There will be an opportunity to provide feedback to the editors if you feel a study shouldn't be reviewed using this model.

Instructions for the Pre-review Phase

You have received a research proposal for pre-review. At this stage, we are asking you to evaluate a proposed research study, looking both at the research value of the study (and its possible results) and at the appropriateness of the methods proposed to carry out the study.

Please remember that this review process is intended to be both evaluative and formative. We are relying on you to exercise judgment about whether the proposed work can reasonably be expected to result in a significant contribution to the Recommender Systems research literature. At the same time, we are also asking you to guide the proposers of the work to methods that would give you (and other experts) confidence in the results and the ability to use them. There are four possible outcomes for a pre-review:

- **Reject.** This should be used whenever the proposed research would not be viewed as having sufficient value to warrant publication, even if it were carried out perfectly with the best possible results. It also should be used if someone proposes a study that is so seriously flawed that it cannot be fixed through improvement in the design.
  For example, work proposing a study to show that using a k-nearest-neighbor collaborative filtering system to achieve comparable user-perceived recommendation quality to an SVD-approximating recommender might well be rejected on the grounds that there are already many papers published showing that result, and that it is not an interesting result. Similarly, work that proposes to show that recommender systems increase overall wellness who proposed to test this using the MovieLens 1M dataset may be rejected on the grounds that it is clear the dataset has no wellness data (and is de-identified). A key message is that it is not the reviewers' responsibility to design a study for researchers who propose an interesting question but lack a close-to-correct method.
- **Revise and Resubmit.** This should be used when the proposed research could produce results that are interesting and valuable to the field, but there are issues in the proposal that need to be fixed to make the study correct. These issues should be substantial (this is not a place to edit writing), but examples might include: an inappropriate experimental design (failure to counter-balance assignments in a within-subjects study), lack of sufficient detail to understand and evaluate a design, lack of needed preliminary work to inform the design (e.g., lack of effect size estimation of power analysis). The narrative of such a review should help the proposer understand what they need to fix, including references to best practices or methods papers or handbooks where appropriate.
- **Conditionally Accept.** This result means that the proposal is accepted, but the resulting research will only be published if the research study and results meet certain conditions. The most common condition will be a "one-way" acceptance for a paper testing a research question or hypothesis that is only interesting in one direction. For example, if someone were to propose to experiment with a recommender system to show that replacing 80% of the ratings with random numbers would not diminish user satisfaction, that might be very interesting and novel if they indeed find that user satisfaction is undiminished. But it would not be publishable if user satisfaction diminished, since that would be consistent with expectations from prior research. If a proposal is conditionally accepted, the Associate Editor will work with the reviewers to come up with a single comprehensive set of conditions for later publication.
- **Accept.** This result means that the research proposal is accepted and the results should be published, whatever they are, as long as the study is successfully completed in accordance with its design. For example, consider a well-designed study to see what balance of LLM-chatbot vs. top-k collaborative filtering interaction users select when given a system that offers users both simultaneously. Reviewers may determine that the result is interesting whatever the balance of usage turns out to be. In this case, the researchers now know that they simply have to conduct the study to have it published.

These guidelines cover both the pre-review and post-review phases.

Instructions for the Post-review Phase

The post-review phase starts when a research study that has received an accept or conditional accept has been completed, the paper is written, and it is time to move towards publication. At this stage, your job as a reviewer is no longer to evaluate the research question or design, but simply whether the research was conducted as agreed, whether the paper as written has all of the elements needed for the work to be adequately documents, and in the case of conditional acceptance, whether the conditions have been met.

At this stage, the possible review outcomes are:

- **Reject.** A paper with a conditional acceptance did not meet the condition in a manner that cannot be remedied. For example, if the results are negative and the condition was to accept only with positive results, it should be rejected. Also, reject is an appropriate result when the researchers did not carry out the study as designed. In both cases, researchers could re-submit to another venue that might find merit in what was actually accomplished in the work.
- **Revisions Required.** A paper that is incomplete or that has remediable issues (that does not require revising the experimentation, but might include changes in writing or analysis). This would be the appropriate review result for a paper that failed to include enough detail in the paper (together with the published protocol) for replication or future meta-analysis).
- **Accept.** This is an indication that the study was conducted according to the published protocol and the resulting paper adequately documents the study and results.

4.5.5.3 Discussion – Challenges to Address in Making this New Model Succeed

In the Registered Reports track, submission approval primarily emphasizes the significance of the research questions and soundness of the research protocol, in alignment with open science principles, rather than the study results. This model is different from the traditional model of publication in most conferences and journals within the recommender system and computer science fields. Therefore, it might require community efforts and time to evolve this initiative to enhance research rigor. By looking at Registered Reports initiatives reflecting in other fields [12], we might foresee challenges in successfully implementing this new model. At the forefront of our considerations are the following:

Enhance awareness of open science in the whole community In our research community, as well as the broader computer science research community, only a few journals or conferences offer Registered Reports. For instance, in software engineering research, Registered Reports were first introduced in 2020 at the International Conference on Mining Software Repositories [19]. It might take some time for researchers to receive the necessary training and education to raise awareness and understanding of open science principles and to understand the benefits of such initiatives. We hope this proposed track can serve as a starting point for the entire community, fostering collective efforts to improve research rigor.

Motivate submission of Registered Reports earlier enough The most obvious barrier to Registered Reports is time. Researchers need to wait for peer review feedback and approval before conducting their study and collecting data, which can be challenging for those on short-term contracts or within short funding cycles. Additionally, students and early-career researchers may need to acquire the necessary skills and knowledge to write a registered report, potentially causing delays in submission and then study

execution. While registered reports could benefit our research in the long run, we also need to consider how to better support researchers at different stages in submitting their registered reports early enough to address these practical challenges.

Ensure effective peer-review Another significant challenge is to ensure that the peer review process functions well, as the benefits of this Registered Reports initiative highly depend on it. This requires both the careful selection of qualified reviewers and successfully engaging them in the review process. Currently, peer reviewers in the community are not always trained to engage deeply with the experimental design and study procedure of the submissions they reviewed. Additional training for reviewers may be necessary for this new model, which might pose further challenges for editors in finding suitable reviewers. To help address this issues, we may consider providing some form of credit for reviewers who offer substantial formative feedback to authors in this model.

References

1 Gediminas Adomavicius, Dietmar Jannach, Stephan Leitner, and Jingjing Zhang. Understanding longitudinal dynamics of recommender systems with agent-based modeling and simulation. *CoRR*, abs/2108.11068, 2021.

2 Icek Ajzen. The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2):179–211, December 1991.

3 Ashton Anderson, Lucas Maystre, Ian Anderson, Rishabh Mehrotra, and Mounia Lalmas. Algorithmic effects on the diversity of consumption on spotify. In *Proceedings of the web conference 2020*, pages 2155–2165, 2020.

4 Ivana Andjelkovic, Parra Denis, and John O'Donovan. Recommendations with optimal combination of feature-based and item-based preferences. In *24th Conference on User Modeling, Adaptation and Personalization (UMAP 2016)*, pages 269–273. ACM Press, 2016.

5 Md Sanzeed Anwar, Grant Schoenebeck, and Paramveer S. Dhillon. Filter bubble or homogenization? disentangling the long-term effects of recommendations on user consumption patterns. In Tat-Seng Chua, Chong-Wah Ngo, Ravi Kumar, Hady W. Lauw, and Roy Ka-Wei Lee, editors, *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 123–134. ACM, 2024.

6 Kirstin C. Appelt, Kerry F. Milch, Michel J. J. Handgraaf, and Elke U. Weber. The decision making individual differences inventory and guidelines for the study of individual differences in judgment and decision-making research. *Judgment and Decision Making*, 6(3):252–262, 2011.

7 Maryam Archie, Sophie Gershon, Abigail Katcoff, and Aileen Zeng. Who's watching? de-anonymization of netflix reviews using amazon reviews. Technical report, MIT, 2018.

8 Guy Aridor, Duarte Goncalves, and Shan Sikdar. Deconstructing the filter bubble: User decision-making and recommender systems. In *Proceedings of the 14th ACM conference on recommender systems*, pages 82–91, 2020.

9 Deanna M Barch. Preregistration and registered reports: A key pathway to enhancing robustness and replicability in mental health research. *Biological Psychiatry: Global Open Science*, 1(2):80–82, 2021.

10 Jordan Barria-Pineda, Kamil Akhuseyinoglu, Stefan Želem Ćelap, Peter Brusilovsky, Aleksandra Klasnja Milicevic, and Mirjana Ivanovic. Explainable recommendations in a personalized programming practice system. In Ido Roll, Danielle McNamara, Sergey Sosnovsky, Rose Luckin, and Vania Dimitrova, editors, *22nd International Conference on Artificial Intelligence in Education, AIED 2021*, volume 12748 of *Lecture Notes in Computer Science*, pages 64–76, Cham, 2021. Springer.

11 R. Bettman, M. Luce, and J. Payne. Constructive consumer choice processes. *Journal of Consumer Research*, 25:187–217, 1998.

12 Christopher D Chambers and Loukia Tzavella. The past, present and future of registered reports. *Nature human behaviour*, 6(1):29–42, 2022.

13 Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM conference on recommender systems*, pages 224–232, 2018.

14 Li Chen, Guanliang Chen, and Feng Wang. Recommender systems based on user reviews: the state of the art. *User Modeling and User-Adapted Interaction*, 25:99–154, 2015.

15 Ine Coppens, Luc Martens, and Toon De Pessemier. Analyzing accuracy versus diversity in a health recommender system for physical activities: a longitudinal user study. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1146–1151, 2023.

16 Pavel Dmitriev, Brian Frasca, Somit Gupta, Ron Kohavi, and Garnet Vaz. Pitfalls of long-term online controlled experiments. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 1367–1376, 2016.

17 Hendrik Drachsler, Katrien Verbert, Olga Santos, and Nikos Manouselis. Panorama of recommender systems to support learning. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 421–451. Springer, Boston, MA, 2015.

18 Michael D. Ekstrand and Martijn C. Willemsen. Behaviorism is not enough: Better recommendations through listening to users. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, page 221–224, New York, NY, USA, 2016. Association for Computing Machinery.

19 Neil A Ernst and Maria Teresa Baldassarre. Registered reports in software engineering. *Empirical Software Engineering*, 28(2):55, 2023.

20 Sandra Garcia Esparza, Michael P O'Mahony, and Barry Smyth. Towards the profiling of twitter users for topic-based filtering. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 273–286. Springer, 2012.

21 Francesco Fabbri, Maria Luisa Croci, Francesco Bonchi, and Carlos Castillo. Exposure inequality in people recommender systems: The long-term effects. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 194–204, 2022.

22 Francesco Fabbri, Yanhao Wang, Francesco Bonchi, Carlos Castillo, and Michael Mathioudakis. Rewiring what-to-watch-next recommendations to reduce radicalization pathways. In *Proceedings of the ACM Web Conference 2022*, pages 2719–2728, 2022.

23 Kim Falk. *Practical recommender systems*. Simon and Schuster, 2019.

24 Sébastien Fassiaux. Preserving consumer autonomy through european union regulation of artificial intelligence: A long-term approach. *European Journal of Risk Regulation*, 2023.

25 A. Felfernig and M.C. Willemsen. *Handling preferences*, pages 91–103. SpringerBriefs in Electrical and Computer Engineering book series. Springer, Germany, 2018.

26 Andres Ferraro, Dietmar Jannach, and Xavier Serra. Exploring longitudinal effects of session-based recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 474–479, 2020.

27 Bruce Ferwerda and Mark P. Graus. Predicting musical sophistication from music listening behaviors: A preliminary study. *CoRR*, abs/1808.07314, 2018.

28 Leon Festinger. *A theory of cognitive dissonance*. A theory of cognitive dissonance. Stanford University Press, 1957. Pages: xi, 291.

29 Sandra Garcia Esparza, Michael P O'Mahony, and Barry Smyth. Catstream: categorising tweets for user profiling and stream filtering. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 25–36, 2013.

30 Erin Gustafson. Meaningful metrics: How data sharpened the focus of product teams, 2023.

31 Tom E Hardwicke and Eric-Jan Wagenmakers. Reducing bias, increasing transparency and calibrating confidence with preregistration. *Nature Human Behaviour*, 7(1):15–26, 2023.

32 J. Hari. *Stolen Focus: Why You Can't Pay Attention*. Bloomsbury Publishing, 2023.

33 Hanna Hauptmann, Nadja Leipold, Mira Madenach, Monika Wintergerst, Martin Lurz, Georg Groh, Markus Böhm, Kurt Gedrich, and Helmut Krcmar. Effects and challenges of using a nutrition assistance system: results of a long-term mixed-method study. *User Modeling and User-Adapted Interaction*, pages 1–53, 2022.

34 A. Jameson, M. Willemsen, A. Felfernig, M. de Gemmis, P. Lops, G. Semeraro, and L. Chen. Human decision making and recommender systems. *Recommender Systems Handbook*, pages 619–655, 2015.

35 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.

36 Dietmar Jannach and Christine Bauer. Escaping the McNamara Fallacy: Towards more impactful recommender systems research. *AI Magazine*, 41(4):79–95, December 2020.

37 Dietmar Jannach and Michael Jugovac. Measuring the business value of recommender systems. *ACM Transactions on Management Information Systems*, 10(4):1–23, December 2019.

38 Dietmar Jannach, Iman Kamehkhosh, and Geoffray Bonnin. *Music Recommendations*, pages 481–518. World Scientific, 2018.

39 Dietmar Jannach, Lukas Lerche, Iman Kamehkhosh, and Michael Jugovac. What recommenders recommend: an analysis of recommendation biases and possible countermeasures. *User Modeling and User-Adapted Interaction*, 25:427–491, 2015.

40 Dietmar Jannach, Lukas Lerche, and Markus Zanker. Recommending based on implicit feedback. In Peter Brusilovsky and Daqing He, editors, *Social Information Access*, volume 10100 of *LNCS*, page in this volume. Springer, Heidelberg, 2017.

41 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.

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

43 Yucheng Jin, Nyi Nyi Htun, Nava Tintarev, and Katrien Verbert. Contextplay: Evaluating user control for context-aware music recommendation. In *the 27th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2019*, pages 294–302. ACM, 2019.

44 Eric J Johnson, Gerald Häubl, and Anat Keinan. Aspects of endowment: a query theory of value construction. *Journal of experimental psychology: Learning, memory, and cognition*, 33(3):461, 2007.

45 Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. News recommender systems – survey and roads ahead. *Information Processing & Management*, 54(6):1203 – 1227, 2018.

46 Poruz Khambatta, Shwetha Mariadassou, Joshua Morris, and S. Christian Wheeler. Tailoring recommendation algorithms to ideal preferences makes users better off. *Scientific Reports*, 13(1):9325, June 2023. Publisher: Nature Publishing Group.

47 Hans K Klein and Daniel Lee Kleinman. The social construction of technology: Structural considerations. *Science, Technology, & Human Values*, 27(1):28–52, 2002.

48 Daniel Kluver, Michael Ekstrand, and Joseph Konstan. Rating-based collaborative filtering: Algorithms and evaluation. In Peter Brusilovsky and Daqing He, editors, *Social Information Access*, LNCS, page 344–390. Springer, Heidelberg, 2017.

49 Bart P Knijnenburg and Martijn C Willemsen. Evaluating recommender systems with user experiments. In *Recommender systems handbook*, pages 309–352. Springer, 2015.

50 Ron Kohavi, Diane Tang, and Ya Xu. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press, 2020.

51 Robert Kraut and Sara Kiesler. Internet paradox revisited. *Journal of Social Issues*, 58(1):49–74, 2002.

52 Su Mon Kywe, Ee-Peng Lim, and Feida Zhu. A survey of recommender systems in twitter. In *Social Informatics: 4th International Conference, SocInfo 2012, Lausanne, Switzerland, December 5-7, 2012. Proceedings 4*, pages 420–433. Springer, 2012.

53 Yu Liang and Martijn C Willemsen. Exploring the longitudinal effects of nudging on users' music genre exploration behavior and listening preferences. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 3–13, 2022.

54 Sebastian Lubos, Alexander Felfernig, and Markus Tautschnig. An overview of video recommender systems: state-of-the-art and research issues. *Frontiers in Big Data, Sec. Recommender Systems*, 6, 2023.

55 M Macleod and D Howells. Protocols for laboratory research. evidence-based preclinical medicine. 2016; 3 (2): e00021, 2016.

56 Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 2145–2148, 2020.

57 Lien Michiels, Jens Leysen, Annelien Smets, and Bart Goethals. What are filter bubbles really? a review of the conceptual and empirical work. In *Adjunct proceedings of the 30th ACM conference on user modeling, adaptation and personalization*, pages 274–279, 2022.

58 Igor H Murai, Alan L Fernandes, Lucas P Sales, Ana J Pinto, Karla F Goessler, Camila SC Duran, Carla BR Silva, André S Franco, Marina B Macedo, Henrique HH Dalmolin, et al. Effect of a single high dose of vitamin d3 on hospital length of stay in patients with moderate to severe covid-19: a randomized clinical trial. *Jama*, 325(11):1053–1060, 2021.

59 Daniel Müllensiefen, Bruno Gingras, Jason Musil, and Lauren Stewart. The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population. *PLoS ONE*, 9(2):e89642, February 2014.

60 Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*, pages 677–686, 2014.

61 Brian A Nosek, Charles R Ebersole, Alexander C DeHaven, and David T Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.

62 Brian A. Nosek and Daniël Lakens. Registered reports: A method to increase the credibility of published reports. *Social Psychology*, 45(3):137–141, 2016.

63 Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, USA, 2016.

64 Derek O'Callaghan, Derek Greene, Maura Conway, Joe Carthy, and Pádraig Cunningham. Down the (white) rabbit hole: The extreme right and online recommender systems. *Social Science Computer Review*, 33(4):459–478, 2015.

65 Eli Pariser. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.

66 Nathalie Percie du Sert, Ian Bamsey, Simon T Bate, Manuel Berdoy, Robin A Clark, Innes Cuthill, Derek Fry, Natasha A Karp, Malcolm Macleod, Lawrence Moon, et al. The experimental design assistant. *PLoS biology*, 15(9):e2003779, 2017.

67 Lorenzo Porcaro, Emilia Gómez, and Carlos Castillo. Assessing the impact of music recommendation diversity on listeners: A longitudinal study. *ACM Transactions on Recommender Systems*, 2(1):1–47, 2024.

68 Lucy Portnoff, Erin Gustafson, Joseph Rollinson, and Klinton Bicknell. Methods for language learning assessment at scale: Duolingo case study. In Sharon I-Han Hsiao,

Shaghayegh (Sherry) Sahebi, François Bouchet, and Jill-Jênn Vie, editors, *Proceedings of the 14th International Conference on Educational Data Mining, EDM 2021, virtual, June 29 - July 2, 2021*. International Educational Data Mining Society, 2021.

69 James O Prochaska and Wayne F Velicer. The transtheoretical model of health behavior change. *American journal of health promotion*, 12(1):38–48, 1997.

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

71 Arianna Sala, Lorenzo Porcaro, and Emilia Gómez. Social media use and adolescents' mental health and well-being: An umbrella review. *Computers in Human Behavior Reports*, 14:100404, 2024.

72 Hanna Schäfer and Martijn C. Willemsen. Rasch-based tailored goals for nutrition assistance systems. In Wai-Tat Fu, Shimei Pan, Oliver Brdiczka, Polo Chau, and Gaelle Calvary, editors, *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI 2019, Marina del Ray, CA, USA, March 17-20, 2019*, pages 18–29. ACM, 2019.

73 Vera Sigre-Leirós, Joël Billieux, Christine Mohr, Pierre Maurage, Daniel L King, Adriano Schimmenti, and Maèva Flayelle. Binge-watching in times of covid-19: A longitudinal examination of changes in affect and tv series consumption patterns during lockdown. *Psychology of Popular Media*, 12(2):173, 2023.

74 Annelien Smets, Jonathan Hendrickx, and Pieter Ballon. We're in this together: a multi-stakeholder approach for news recommenders. *Digital Journalism*, 10(10):1813–1831, 2022.

75 Alain Starke, Martijn C. Willemsen, and Chris Snijders. Effective user interface designs to increase energy-efficient behavior in a rasch-based energy recommender system. In Paolo Cremonesi, Francesco Ricci, Shlomo Berkovsky, and Alexander Tuzhilin, editors, *Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys 2017, Como, Italy, August 27-31, 2017*, pages 65–73. ACM, 2017.

76 Alain D. Starke, Martijn C. Willemsen, and Chris Snijders. Using explanations as energy-saving frames: A user-centric recommender study. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, pages 229–237. ACM, 2021.

77 Galen Stocking, Amy Mitchell, Katerina Eva Matsa, Regina Widjaya, Mark Jurkowitz, Shreenita Ghosh, Aaron Smith, Sarah Naseer, and Christopher St Aubin. The role of alternative social media in the news and information environment. *Pew Research Center*, 2022.

78 Marko Tkalčič, Matevž Kunaver, and Jurij Tasic. Personality Based User Similarity Measure for a Collaborative Recommender System. In *Proceedings of the 5th Workshop on Emotion in Human-Computer Interaction – Real world challenges*. Fraunhofer Verlag, 2009.

79 Helma Torkamaan. Mood measurement on smartphones: Which measure, which design? *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 7(1), mar 2023.

80 Helma Torkamaan, Catalin-Mihai Barbu, and Jürgen Ziegler. How can they know that?: a study of factors affecting the creepiness of recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems – RecSys '19*, pages 423–427. ACM Press, 2019.

81 Helma Torkamaan and Jürgen Ziegler. Mobile mood tracking: An investigation of concise and adaptive measurement instruments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(4), dec 2020.

82 Helma Torkamaan and Jürgen Ziegler. Recommendations as challenges: Estimating required effort and user ability for health behavior change recommendations. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, IUI '22, page 106–119, New York, NY, USA, 2022. Association for Computing Machinery.

83 Helma Torkamaan and Jürgen Ziegler. A taxonomy of mood research and its applications in computer science. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 421–426, 2017. ISSN: 2156-8111.

84 Helma Torkamaan and Jürgen Ziegler. Exploring chatbot user interfaces for mood measurement: a study of validity and user experience. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, UbiComp-ISWC '20, pages 135–138. Association for Computing Machinery, 2020.

85 John C. Turner and Penelope J. Oakes. The significance of the social identity concept for social psychology with reference to individualism, interactionism and social influence. *British Journal of Social Psychology*, 25(3):237–252, 1986.

86 UNESCO. Recommendation on the ethics of artificial intelligence. Technical Report SHS/BIO/REC-AIETHICS/2021, UNESCO, 2021.

87 Ibo van de Poel. Embedding values in Artificial Intelligence (AI) systems. *Minds and Machines*, 30(3):385–409, September 2020.

88 Jingjing Zhang, Gediminas Adomavicius, Alok Gupta, and Wolfgang Ketter. Consumption and performance: Understanding longitudinal dynamics of recommender systems via an agent-based simulation framework. *Information Systems Research*, 31(1):76–101, 2020.

89 Meizi Zhou, Jingjing Zhang, and Gediminas Adomavicius. Longitudinal impact of preference biases on recommender systems' performance. *Information Systems Research*, 2023.

Participants

- Gediminas Adomavicius
University of Minnesota – Minneapolis, US
- Vito Walter Anelli
Politecnico di Bari, IT
- Andrea Barraza-Urbina
Grubhub – New York, US
- Christine Bauer
Paris Lodron University Salzburg – AT
- Joeran Beel
Universität Siegen, DE
- Alejandro Bellogín
Autonomous University of Madrid, ES
- Toine Bogers
IT University of Copenhagen, DK
- Peter Brusilovsky
University of Pittsburgh, US
- Robin Burke
University of Colorado – Boulder, US
- Wanling Cai
Lero, the Science Foundation Ireland – Limerick, IE
& Trinity College – Dublin, IE
- Tommaso Di Noia
Politecnico di Bari, IT
- Michael D. Ekstrand
Drexel University – Philadelphia, US
- Kim Falk
DPG Media – Antwerp, BE
- Andres Ferraro
SiriusXM, US
- Bart Goethals
University of Antwerp, BE
- Neil Hurley
University College Dublin, IE
- Dietmar Jannach
University of Klagenfurt, AT
- Olivier Jeunen
ShareChat – Edinburgh, GB
- Joseph Konstan
University of Minnesota – Minneapolis, US
- Dominik Kowald
Know Center – Graz, AT
& TU Graz, AT
- Maria Maistro
University of Copenhagen, DK
- Lien Michiels
University of Antwerp, BE
- Julia Neidhardt
TU Wien, AT
- Özlem Özgöbek
NTNU – Trondheim, NO
- Denis Parra
PUC – Santiago de Chile, CL
- Sole Pera
TU Delft, NL
- Lorenzo Porcaro
EC Joint Research Centre – Ispra, IT
- Alan Said
University of Gothenburg, SE
- Rodrygo Santos
Federal University of Minas Gerais-Belo Horizonte, BR
- Guy Shani
Ben Gurion University – Beer Sheva, IL
- Manel Slokom
TU Delft, NL
- Annelien Smets
Vrije Universiteit Brussel, BE
- Barry Smyth
University College Dublin, IE
- Marko Tkalcic
University of Primorska, SI
- Helma Torkamaan
TU Delft, NL
- Alexander Tuzhilin
New York University, US
- Tobias Vente
Universität Siegen, DE
- Robin Verachtert
DPG Media – Antwerp, BE
- Lukas Wegmeth
Universität Siegen, DE
- Martijn Willemsen
TU Eindhoven, NL
- Eva Zangerle
University of Innsbruck, AT
- Jürgen Ziegler
Universität Duisburg-Essen, DE