The Quest for Reliable Metrics of Responsible AI

Theresia Veronika Rampisela *University of Copenhagen* thra@di.ku.dk
Maria Maistro *University of Copenhagen* mm@di.ku.dk
Tuukka Ruotsalo *University of Copenhagen and LUT University* tr@di.ku.dk
Christina Lioma *University of Copenhagen* c.lioma@di.ku.dk

Abstract

The development of Artificial Intelligence (AI), including AI in Science (AIS), should be done following the principles of responsible AI. Progress in responsible AI is often quantified through evaluation metrics, yet there has been less work on assessing the robustness and reliability of the metrics themselves. We reflect on prior work that examines the robustness of fairness metrics for recommender systems as a type of AI application and summarise their key takeaways into a set of non-exhaustive guidelines for developing reliable metrics of responsible AI. Our guidelines apply to a broad spectrum of AI applications, including AIS.

Keywords: responsible AI, fairness evaluation, recommender systems

1 Introduction

Recent legislation, such as the EU Artificial Intelligence (AI) Act and the EU Digital Services Act, has increasingly emphasised the responsible development of AI applications [European Parliament and Council, 2022, 2024] to protect their users and minimise harms. One such AI application is Recommender Systems (RSs). RSs are highly useful in finding relevant items that match user needs, preferences, or interests [Aggarwal, 2016]. These systems are widely used for both high- and low-stakes activities in professional and personal settings. In professional settings, among other use cases, RSs can be used to suggest scientific papers and citations for researchers [Kreutz and Schenkel, 2022, Färber and Jatowt, 2020], or job positions for jobseekers [Siting et al., 2012].

Prior work has investigated aspects of responsible AI in RSs, such as fairness and bias [Wang et al., 2023, Klimashevskaia et al., 2024]. Biases may exist in the data, the algorithm, or the evaluation pipeline of an RS; all these biases can contribute to unfairness. While there is no universal definition, fairness in RSs is commonly understood as providing equal treatment or equitable outcomes to the RS stakeholders [Wang et al., 2023], ensuring that no individuals or groups are systematically discriminated.

Unfair RSs can have severe real-life consequences. In job recommendations, for example, unfair RSs may contribute to the exacerbation of gender pay gaps. This could happen if historically marginalised groups (e.g., women) are only shown recommendations for lower-paying jobs, while highly-paid positions are

only recommended to the historically dominant group. In the AIS context, an unfair paper/citation RS could perform extremely well in recommending relevant work for one discipline (e.g., computer science) but perform poorly for other disciplines (e.g., Nordic studies) due to training data imbalance between various fields of study, hindering the development of sciences. Similarly, if an unfair RS overpromotes articles by researchers from economically developed countries [Biega et al., 2020], it may simultaneously provide less exposure to researchers from other countries. This imbalance could lead to a less inclusive understanding of science [Mihalcea et al., 2025], especially in social sciences and humanities, where cultural context in particular matters. Hence, the development of fair RSs is not only scientifically relevant but also societally important.

Many evaluation approaches and metrics have been proposed to measure RS fairness. Yet, few studies have investigated the reliability and robustness of these approaches. This leads to the metric score being potentially misleading or unstable, and thus unreliable. Having reliable metrics is critical as they primarily guide the responsible development of AI(S) applications, especially in the early stages, as they provide a scalable way of evaluating system performance without significant cost (e.g., opposed to user studies). In this paper, we condense insights from our previous work on RS fairness evaluation [Rampisela et al., 2024a,b, 2025a,b] into a set of guidelines for developing reliable metrics of responsible AI, based on the limitations that we uncovered in existing approaches. As these guidelines are general, they would also be relevant for evaluating AI in science (AIS).

2 Fairness Evaluation

In this part, we explain how fairness, an aspect of responsible AI, is evaluated in RSs (§2.1). We summarise our findings and contributions from previous work on RS fairness evaluation (§2.2).

2.1 Evaluation of Fairness in Recommender Systems

At a high level, Recommender System (RS) fairness evaluation is often categorised in terms of the *subject* (user/item fairness) and the evaluation *granularity* (group/individual fairness) [Wang et al., 2023]. *User fairness* concerns the disparity in the recommendation effectiveness, where an RS that performs equally well for all users would be deemed fairer than an RS that can perform very well for some users but fails to provide relevant recommendations for others. In contrast, *item fairness* is usually defined in terms of how much exposure is received by the item, in comparison to other items in the RS. In terms of granularity, *group fairness* concerns the difference in utility (recommendation effectiveness or item exposure) between subject groups, where the groups are typically formed based on user or item attributes, e.g., socio-demographic attributes. Meanwhile, *individual fairness* is frequently operationalised as the variation of utility across all users or items.

RS practitioners often rely on computing metrics as a proxy for fairness. Over 30 metrics have been used to measure fairness for different subjects and at different granularity levels [Wang et al., 2023]. Some metrics are based on well-known inequality measures, such as standard deviation, entropy, and Gini Index. Others are newly developed based on existing fairness notions, e.g., Rawlsian fairness and envy-freeness. With the sheer amount of metrics, it is often difficult to understand what and how they actually quantify fairness. Even worse, some metrics were proposed without further analysis of their limitations. Consequently, the measure limitations are largely unknown.

2.2 Our Contributions to RS Fairness Evaluation

We analysed several RS fairness metrics and uncovered limitations within the measures. To resolve the limitations, we proposed novel fairness evaluation approaches. Finally, we provided practical guidelines for evaluating RS fairness.

Analysis on metric limitations. Our investigation on metric limitations was done both theoretically and empirically. In our theoretical studies, we show that some fairness metrics are mathematically flawed [Rampisela et al., 2024a, 2025a]. Firstly, some metrics would either crash upon computation due to invalid mathematical operations (e.g., division by 0), resulting in the metric outputting no scores that can be used for fairness assessment, rendering it unusable. Secondly, a large number of the metrics have an unknown score range, or if it is known, the max/min score is not reachable. Suppose that in theory, a metric score ranges between 0 and 1, where 0 is the fairest score and 1 is the unfairest score. In some cases, there is no input to the metric that would output 0 or 1. Instead, for example, its score could only range between 0.3 and 0.6 (i.e., the minimum and maximum reachable scores, respectively). This causes difficulty in interpreting the metric score, as one would interpret a score of 0.5 as somehow fair if the range is $[0, 1]$, but unfair if the range is $[0.3, 0.6]$. Thirdly, in some cases, it is not even known what kind of input to the metric would result in the max/min scores, resulting in a limited understanding of what the (un)fairest possible case is, according to the metric.

Our empirical studies showed that some metrics tend to score very low (close to 0) independently of the fairness level. The compressed score range and limited metric sensitivity give the illusion of an extremely fair input, even when it is not fair, misleading the understanding of how (un)fair a system is. Additionally, we found that some metrics may be redundant as they yield similar conclusions to other existing metrics. As such, computing one of the similar metrics may suffice. On the other hand, we showed that metrics for one granularity level cannot be used as a proxy to the other, i.e., group fairness metrics cannot estimate individual fairness [Rampisela et al., 2025b]. Therefore, it is necessary to consider different granularities to have a more comprehensive view of fairness.

By investigating the metrics' limitations, we hope to raise awareness that a metric could suffer from several issues that could affect its usability and in-

terpretability. Thus, one should not immediately trust a metric's score at face value without truly understanding how the metric operates.

New evaluation approaches. We contributed two types of new evaluation approaches: metric corrections and new metrics. Firstly, we correct existing fairness metrics by redefining their formulation to avoid computation crash due to invalid mathematical operations. We also applied min-max normalisation, such that the score 0 is correctly mapped to the fairest possible case and 1 to the unfairest possible case [Rampisela et al., 2024a, 2025a]. The corrected metrics can then be interpreted more easily than the original metrics. Secondly, we proposed a new metric to jointly evaluate recommender system effectiveness and fairness, as existing metrics cannot quantify both aspects simultaneously [Rampisela et al., 2025c]. We have also released the source codes to compute both new evaluation approaches publicly, so that they can be used more easily.[1]

Practical guidelines. Considering that there are many RS fairness metrics with their own limitations, we summarise practical guidelines for selecting the metrics from the findings in our previous work. Firstly, we recommend using our corrected metrics to evaluate how close a recommendation is to the fairest recommendation scenario, as the original metrics cannot be used to do so. Secondly, the metric scores should be interpreted carefully, as some measures tend to score very close to the fairest case and may overestimate fairness. Thirdly, the use of redundant (highly similar) measures should be avoided. Fourthly, one should evaluate for both group and individual fairness to ensure that no groups or individuals are disadvantaged.

3 Guidelines for Formulating Reliable Metrics

With the rise of AI usage in science, it is inevitable that new metrics may be formulated to evaluate their performance, including how well they align with responsible AI aspects. Reflecting and generalising on insights from our work on RS fairness evaluation, we provide a set of questions to guide the development of reliable metrics for Responsible AI:

1. Are there input cases that should be excluded, such that the metric does not have invalid mathematical operations?
2. What is the metric range and how should it be interpreted?
3. What kind of input results in the minimum and maximum metric score?
4. How sensitive is the metric to changes in the input?
5. Does the metric yield a similar conclusion to an existing one?

[1] For the corrected metrics, please refer to github.com/theresiavr/individual-item-fairness-measures-recsys and github.com/theresiavr/relevance-aware-item-fairness-measures-recsys. For the effectiveness-fairness joint evaluation approach, please refer to github.com/theresiavr/DPFR-recsys-evaluation.

4 Conclusions

Our guidelines are not meant to be exhaustive. Rather, these are the bare minimum to ensure reliable metrics for AI technologies. We intend to raise awareness, so that effort should be put not only into developing new technology, but also into improving its evaluation. Quantification is an important part of regulation and policy-making, yet evaluation metrics are frequently absent from the current AI policy discussion.[2] As such, future work could potentially involve collaboration with AI users, technical actors, social scientists, experts in ethics, policy makers, and governmental agencies to discuss the appropriate Responsible AI evaluation metric(s) to incorporate in AI-related policies and regulations, especially in high-stake contexts such as AIS. The collaboration could potentially study which responsible AI aspects are critical for AIS and perform an audit of existing evaluation metrics to ensure their reliability in measuring the intended aspects. We expect the collaboration to provide more concrete, measurable guidelines for the responsible development of AI applications.

References

C. C. Aggarwal. *Recommender Systems: The Textbook*. Springer Publishing Company, Incorporated, 1st edition, 2016. ISBN 3319296574.

A. J. Biega, F. Diaz, M. D. Ekstrand, S. Feldman, and S. Kohlmeier. Overview of the TREC 2020 Fair Ranking Track. In *The Twenty-Eighth Text REtrieval Conference (TREC 2020) Proceedings*, 2020.

European Parliament and Council. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance), 2022. URL `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022R2065`.

European Parliament and Council. Regulation (EU) 2024/1689: Artificial Intelligence Act, 2024. URL `https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng`. Recital 27 discusses principles for trustworthy and ethical AI.

M. Färber and A. Jatowt. Citation recommendation: approaches and datasets. *Int. J. Digit. Libr.*, 21(4):375–405, Dec. 2020. ISSN 1432-5012. doi: 10.1007/s00799-020-00288-2. URL `https://doi.org/10.1007/s00799-020-00288-2`.

A. Klimashevskaia, D. Jannach, M. Elahi, and C. Trattner. A survey on popularity bias in recommender systems. *User Modeling and User-Adapted Interaction*, 34(5):1777–1834, July 2024. ISSN 0924-1868. doi: 10.1007/s11257-024-09406-0. URL `https://doi.org/10.1007/s11257-024-09406-0`.

[2] `https://oecd.ai/en/catalogue/overview`

C. K. Kreutz and R. Schenkel. Scientific paper recommendation systems: a literature review of recent publications. *Int. J. Digit. Libr.*, 23(4):335–369, Dec. 2022. ISSN 1432-5012. doi: 10.1007/s00799-022-00339-w. URL https://doi.org/10.1007/s00799-022-00339-w.

R. Mihalcea, O. Ignat, L. Bai, A. Borah, L. Chiruzzo, Z. Jin, C. Kwizera, J. Nwatu, S. Poria, and T. Solorio. Why ai is weird and shouldn't be this way: towards ai for everyone, with everyone, by everyone. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i27.35092. URL https://doi.org/10.1609/aaai.v39i27.35092.

T. V. Rampisela, M. Maistro, T. Ruotsalo, and C. Lioma. Evaluation Measures of Individual Item Fairness for Recommender Systems: A Critical Study. *ACM Trans. Recomm. Syst.*, 3(2), 11 2024a. doi: 10.1145/3631943. URL https://doi.org/10.1145/3631943.

T. V. Rampisela, T. Ruotsalo, M. Maistro, and C. Lioma. Can We Trust Recommender System Fairness Evaluation? The Role of Fairness and Relevance. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pages 271–281, New York, NY, USA, 2024b. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657832. URL https://doi.org/10.1145/3626772.3657832.

T. V. Rampisela, M. Maistro, T. Ruotsalo, F. Scholer, and C. Lioma. Relevance-aware individual item fairness measures for recommender systems: Limitations and usage guidelines. *ACM Trans. Recomm. Syst.*, Sept. 2025a. doi: 10.1145/3765624. URL https://doi.org/10.1145/3765624. Just Accepted.

T. V. Rampisela, M. Maistro, T. Ruotsalo, F. Scholer, and C. Lioma. Stairway to fairness: Connecting group and individual fairness. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, RecSys '25, page 677–683, New York, NY, USA, 2025b. Association for Computing Machinery. ISBN 9798400713644. doi: 10.1145/3705328.3748031. URL https://doi.org/10.1145/3705328.3748031.

T. V. Rampisela, T. Ruotsalo, M. Maistro, and C. Lioma. Joint evaluation of fairness and relevance in recommender systems with pareto frontier. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 1548–1566, New York, NY, USA, 2025c. Association for Computing Machinery. ISBN 9798400712746. doi: 10.1145/3696410.3714589. URL https://doi.org/10.1145/3696410.3714589.

Z. Siting, H. Wenxing, Z. Ning, and Y. Fan. Job recommender systems: A survey. In *2012 7th International Conference on Computer Science & Education (ICCSE)*, pages 920–924, 2012. doi: 10.1109/ICCSE.2012.6295216.

Y. Wang, W. Ma, M. Zhang, Y. Liu, and S. Ma. A Survey on the Fairness of Recommender Systems. *ACM Trans. Inf. Syst.*, 41(3), 2 2023. ISSN 1046-8188. doi: 10.1145/3547333. URL `https://doi.org/10.1145/3547333`.