Toward Operationalizing a Comprehensive Evaluation Framework for Recommender Systems Explanations

Kathrin Wardatzky[1], Oana Inel[1] and Abraham Bernstein[1]

[1]University of Zurich, Switzerland

Abstract

The evaluation of recommender systems' explanations is an ongoing research challenge with frequent publication of new evaluation approaches and metrics. We noticed differences in evaluation approaches in terms of the aspects of the explanations that are evaluated between research developing new explanation methods or explainable recommender systems (i.e., algorithm-focused studies), and research focusing on the impact of explanations on the users (i.e., user-centered studies). While the algorithmic side focuses on developing new metrics that allow for an offline evaluation and comparison of explanation approaches and their direct output, the user-centered side evaluates formatted explanations that are often simulated and not generated by an explainer. It is rare that an explanation method is evaluated and compared to other approaches from generation to interaction with the user. In this position paper, we argue that a comprehensive evaluation requires a combination of both sides and, therefore, a multidisciplinary effort to develop a feasible approach that captures the strengths and weaknesses of an explanation method. We take the first steps towards moving in this direction by employing a theoretical, comprehensive evaluation framework and discussing the challenges that we came across in this process.

Keywords

Explainable Recommender Systems, Evaluation, Human-Centered AI, Evaluation Metrics

1. Introduction

Recommender systems (RS) are indispensable in our everyday online interactions. They influence what news articles we read to inform ourselves [1, 2], the media we consume for entertainment [3, 4], what items we buy [5], and what jobs we apply for [6]. Explanations are often seen as a means to increase the transparency of RS, help users make better and more informed decisions, and increase their trust and satisfaction [7, 8]. However, there are still no established guidelines or standards for the evaluation of the explanations [9]. This makes it difficult to decide which explanation approach works best for a given RS application. What further increases the challenge to evaluate explanations is the fact that their effect might differ between user types or application domains [10, 11].

Miller [12] divided the process of human-to-human explanation into three elements: (1) the *cognitive process* in which the causes for an event that is to be explained are identified and selected, (2) the *product* of the cognitive process which is the explanation, and (3) the *social process* of transferring the knowledge to the explainee — ideally in a way that they understand the causes of the event. Donoso-Guzmán et al. [13] adapted this definition to AI explanations. The identification and selection of the causes that will be part of the explanation happen in the *generation* stage. The output, or product, of an explainer is often a structured *abstraction* that requires an additional step to design the explanation in the desired *format*. The final social process then happens in the *communication* stage. Thus, multiple points throughout this process require an evaluation of different properties to ensure that the final product is of high quality.

In the context of reading 1,012 papers for a survey on explanations in RS [11], we notice that many publications that propose new explainable RS or explanation methods mainly evaluate the first two elements of the explanation process, without considering the final product and communication. Conversely, the human-computer-interaction-centered papers focus on evaluating the final two elements, often with simulated recommendations and explanations. *We argue that a thorough evaluation requires*

Beyond Algorithms: Reclaiming the Interdisciplinary Roots of Recommender Systems Workshop (BEYOND 2025), September 26th, 2025, co-located with the 19th ACM Recommender Systems Conference, Prague, Czech Republic.

wardatzky@ifi.uzh.ch (K. Wardatzky); inel@ifi.uzh.ch (O. Inel); bernstein@ifi.uzh.ch (A. Bernstein)

0000-0002-7043-7326 (K. Wardatzky); 0000-0003-4691-6586 (O. Inel); 0000-0002-0128-4602 (A. Bernstein)

a holistic view integrating all four elements: generation, abstraction, formatting, and communication. To that end, *we discuss our efforts to build the infrastructure that allows the systematic evaluation and comparison of explanation methods for RS whilst considering all four elements.* We concentrate our work on the evaluation of explanations that will ultimately be shown to an RS end-user. As a foundation, we selected the framework for user-centered evaluation by Donoso-Guzmán et al. [13], as it is built upon the well-known user experience evaluation framework for RS by Knijnenburg et al. [14] and, to the best of our knowledge, the first framework to integrate the dimension of the explanation process. Since the framework of Donoso-Guzmán et al. [13] is intended to evaluate AI explanations in general, in this paper, we map the suggested explanation properties to the recommendation scenario. Our ultimate goal is to build a tool that allows us to evaluate and compare the strengths and weaknesses of explanation methods across multiple dimensions, from explanation generation to the impact on users.

2. Related Work

We now discuss prior work on the evaluation of AI explanations in RS applications and explainable AI (XAI) evaluation in general. For brevity, we limit this section to related work with a focus on the evaluation of explanations. We divide the section into practical and theoretical evaluation frameworks.

2.1. Practical Evaluation Frameworks

With the persistent challenge of evaluating AI explanations, several frameworks aiming to compare different explanation methods were published. These frameworks range from general XAI applications [10, 15] to specific explanation methods (e.g., Agarwal et al. [16] for attribute-based explanations). In the realm of RS, Coba et al. [17] were, to the best of our knowledge, the first to implement a framework focusing on explanations and their evaluation. What these frameworks have in common is that they all focus the evaluation on the generation and abstraction elements of the explanations. One exception we found is Ariza-Casabona et al. [18], who recently published their comparative evaluation approach for text-based explanations in RS, including, thus, the explanation format assessment.

Despite the availability of these frameworks, we still lack a structured evaluation approach that includes all four elements of explanations and that can be applied to multiple XAI approaches in RS, which is why we opted to operationalize a theoretical framework.

2.2. Theoretical Evaluation Frameworks

The challenges of evaluating AI explanations have been pointed out in the literature for many years. Doshi-Velez and Kim [19] proposed one of the first taxonomies of evaluation approaches and discussed the issues connected with each approach. Their taxonomy consists of three evaluation approaches: (1) application-grounded evaluation, where actual users solve the actual task, (2) human-grounded metrics, where users perform a simplified task, and (3) functionally-grounded evaluation, where a formal definition of explanation quality is used as a proxy to evaluate the explanations without users. They point out that the evaluation types inform each other: the functional proxies reflect real-world performance, and the simplified tasks capture the essence of the actual application setting.

Since then, multiple surveys [20, 21, 22] summarized the evaluation approaches for explanations. Some of these surveys resulted in theoretical frameworks and guidelines aiming for a more structured and standardized evaluation approach. Nauta et al. [23] surveyed 361 papers and derived 12 properties of explanation quality. Then, they categorized the quantitative evaluation methods from these papers along the properties to provide suitable objective measures to evaluate and compare the aspects of explanation approaches. In contrast, Donoso-Guzmán et al. [13] surveyed evaluation approaches from a user experience perspective. They categorized the evaluation approaches of 29 papers along the user-centered evaluation framework for RS developed by Knijnenburg et al. [14] and the elements of explanations inspired by Miller [12]. To our knowledge, such a theory-grounded framework that

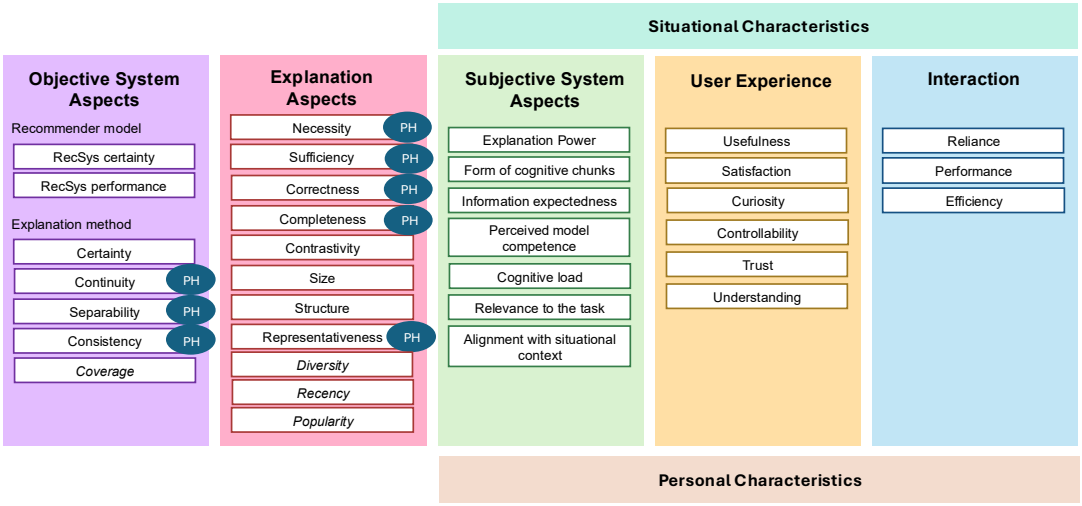

Figure 1: Framework architecture based on Donoso-Guzmán et al. [13]. We extended the **Objective System Aspects** and **Explanation Aspects** with the *italicized* properties, further highlighted the properties with measures focusing on post-hoc explainers with "PH".

combines all elements of an AI explanation and is targeted at RS does not exist. Therefore, we selected this framework as our foundation because it extends an evaluation framework developed for RS.

3. Explanation Properties and Their Measures

In this section, we go through each explanation component of the framework of Donoso-Guzmán et al. [13], map the explanation properties that are evaluated to the RS application, and, if needed, extend the component with properties from the explainable RS literature. We structured the section along the five core elements of the theoretical evaluation framework: the *objective system aspects* describe the system's abilities, the *explanation aspects* group properties describing the quality of an explanation, the *subjective system aspects* include the users' perception of the objective system aspects, the *user experience* describes what the users encounter when interacting with the system, and the *interaction* aspects describe factors relating to a possible system adoption [13, 14]. Additionally, we discuss the importance of considering *situational* and *personal characteristics* dimensions that were not analyzed in the framework due to a lack of information on their impact on the explanation effects. The idea is to evaluate explanation approaches on all components to determine their strengths and weaknesses, to allow AI practitioners or researchers to select the appropriate method for their application and carefully consider possible trade-offs. Figure 1 provides an overview of the explanation properties and their categorization. For brevity, we refer to Table 1 in Donoso-Guzmán et al. [13] for their definitions. Each subsequent section provides an overview of what is already there and then addresses remaining challenges.

3.1. Objective System Aspects

The evaluation of objective system aspects is centered around the explanation generation and abstraction stages and consists of six properties: AI model performance, AI model certainty, XAI certainty, continuity, separability, and consistency. The RS performance and certainty can impact the explanation quality and perception of the overall system [24]. Therefore, the ranking or rating capabilities of the RS should be reported along with the model certainty as part of a comprehensive evaluation of the explanations. There are plenty of evaluation frameworks for RS that offer a variety of metrics to evaluate the recommendation capabilities.[1] On the explanation method side, the certainty property can be evaluated with a checkbox stating whether (un-)certainty values are present in the explanation.

[1]To start with, see the list provided by the RecSys conference: https://github.com/ACMRecSys/recsys-evaluation-frameworks

Adapting metrics from classification to local explanations in top-n recommendations. A series of metrics has been proposed to evaluate the remaining explanation properties (consistency, continuity, and separability) that cover objective system aspects. Both the OpenXAI [16] and Quantus [15] frameworks, for example, compiled a list of consistency and continuity metrics for classification approaches. When applying these consistency and continuity metrics to a top-n recommendation problem with local explanations, the evaluation time, however, increases significantly. In addition, many metrics in the classification domain are designed for post-hoc, feature-importance explanation methods. It is difficult to implement measures that would allow to compare different explanation methods in these properties. One could argue that explanations derived from intrinsically explainable models should be inherently consistent with high degrees of continuity and separability. Having the evaluation results for intrinsically explainable models and being able to compare both approaches could provide a better intuition on how good is, e.g., the continuity of post-hoc explainers.

Need for additional evaluation properties. In addition to the explanation properties mentioned by Donoso-Guzmán et al. [13], we suggest evaluating the *coverage* of the explainer, as in the number of recommendations that can(not) be explained. This property differs from the completeness property, as the latter assumes an explanation exists. Depending on the definition of explainable that is implemented in the model, it might not be possible to explain each recommendation of a top-n list. An example would be the *Novel and Explainable Matrix Factorization* model [25] which follows the intuition that an item is explainable if n users in the neighborhood of a target user have also interacted with the item. Otherwise, no explanation can be provided. The authors suggest evaluating this with a metric derived from information retrieval, Explainable NDCG, which not only measures how many items in a top-n recommendation list can be explained but also their position in the ranking and the ratings of the nearest neighbors. The metric, however, only works for collaborative filtering explanations and requires explicit rating data. A measure that is independent of the explanation approach would be to simply report the average ratio of items without explanation in a top-n recommendation list.

In summary, the main challenge of evaluating objective system aspects is the adaptation of existing metrics from a classification to a recommendation problem.

3.2. Explanation Aspects

The explanation aspects consist of eight properties in the literature: necessity, sufficiency, correctness, completeness, contrastivity, size, structure, and representativeness. All of them are evaluated in the abstraction stage. Additionally, correctness offers evaluation opportunities in the generation and format stages, completeness in the format stage, and structure in the communication stage.

Adapting metrics from classification to local explanations in top-n recommendations. Similar to the challenges mentioned in Section 3.1, the majority of the metrics that are used to evaluate necessity, sufficiency, correctness, completeness, and contrastivity were developed with a classification problem in mind. They are often specific to post-hoc explainers, and one can argue whether these properties should indeed be evaluated for intrinsically explainable models.

Modifying explanation abstraction. In our implementation efforts, we currently treat the structure of an explanation at the abstraction level as a parameter that determines how the explanation will be formatted and presented at the communication level. One could follow, for example, the data visualization guidelines from Chatti et al. [26] to determine a suitable format. A remaining challenge is that there are still too many options that are suitable for most explainer types, even when applying such guidelines. More research is needed to determine the best way to present the explanations for different user types or in different contexts (e.g., application domains, mobile vs. desktop screen) [13, 11].

Dependency of explanation properties to RS. The size of the explanation can be dependent on the RS (e.g., an extracted path of the random walk recommender RP$^3\beta$ [27] will always consist of three

hops). Alternatively, a parameter can determine how many aspects or features should be extracted in the generation process. In all other cases, the average number of elements in an explanation should be reported along with the standard deviation to get an impression of the explanation size.

Additional evaluation properties. On the RS side, there are indications that users prefer to see *diverse explanations* that mention *popular* features and that the explanations are based on the target user's *recent* interaction history [28]. The corresponding metrics have been mostly applied to assess path-based explanations derived from knowledge graphs [29] or feature-based explanations [18, 30], but could be adapted to evaluate the abstraction level of other explanation types. In these use cases the evaluated notion of diversity is often related to the type of nodes and relations that are mentioned in a path-based explanation, or the diversity in the mentioned item features in feature-based explanations. Future work should investigate whether this operationalization of diversity in explanations should be extended to other notions. For brevity, we refer to Vrijenhoek et al. [31] and Heitz et al. [2] for the conceptualization and operationalization of different recommendations diversity notions.

In summary, the evaluation of explanation aspects faces additional challenges besides the metric adaptation (Section 3.1), such as (1) many properties lack the option to compare intrinsically explainable methods with post-hoc approaches, and (2) there are too many options in which an abstract explanation can be formatted. Evaluating all of them, particularly with a user study, is not feasible.

3.3. Subjective System Aspects

The seven properties that fall into the subjective system aspects (explanation power, form of cognitive chunks, information expectedness, perceived model competence, cognitive load, relevance to the tasks, and alignment with situational context) are used to evaluate a user's perception of the explanations. Therefore, all of them are evaluated at the communication stage with users.

Metrics specific to explanation type. To limit the number of dimensions that need to be evaluated with a user study, some metrics have been proposed to evaluate the explanation power, form of cognitive chunks, information expectedness, and relevance to the tasks at the abstraction or format stage. These metrics are either specific to an explanation type, e.g., pragmatism to evaluate the relevance to the task property for counterfactual explanations [23], or the metrics require a ground truth to be compared with, e.g., BLEU and METEOR scores [13], to evaluate the form of cognitive chunks.

Ground truth proxy. In the RS domain, we often see user reviews being used as a proxy for a ground truth for text explanations (see, e.g., Ariza-Casabona et al. [18]). Aside from only being applicable to RS leveraging user reviews in their recommendation process and explanations being presented in natural language, it also raises other challenges related to the nature of user reviews. Reviews are often biased towards very good or very bad opinions as users rarely leave a review if the product or experience met, but did not exceed their expectations [32]. They additionally face the issue that certain user types are more likely to leave a review than others [33].

The main challenge of evaluating the subjective system aspects is a continuation of the issues pointed out in Section 3.2. The proposed metrics to reduce the number of dimensions that would need to be evaluated with a user study are currently not sufficient.

3.4. User Experience and Interaction

All explanation properties in the user experience and interaction are evaluated at the communication stage with users.

Alignment of evaluation goal with measurement. While the interaction properties (reliance, performance, and efficiency) can be captured with implicit measures, such as the number of clicks, time spent in the application, and whether the item that the user selected aligns with a prior goal (i.e.,

selecting healthier meals), the user experience properties often require the users to fill in questionnaires. These questionnaires need to be carefully crafted and consider the notion of the properties that should be evaluated. Taking the understanding property as an example, the results might differ if the user is asked to rate how much a presented explanation increases their understanding of the system, compared to asking questions that capture their understanding before and after seeing the explanations. In these cases, leveraging insights and instruments from social science would benefit the design of the user studies instead of having the user rate how well an explanation is perceived in each property.

The main challenge with the user evaluation of the communication stage is to balance the coverage of properties that should be evaluated and the number of dimensions in the user study.

3.5. Situational and Personal Characteristics

The situational and personal characteristics were not specifically integrated in the framework of Donoso-Guzmán et al. [13] due to a lack of information on how these characteristics impact the explanation properties.

No general knowledge about impact. The impact of individual personal and situational characteristics on explanation properties concerning the user experience and interaction has been explored for specific settings (e.g., [34, 35, 36, 37]), but it is challenging to derive general conclusions from them [11].

Lack of diverse participants. An additional issue is the fact that the majority of the user evaluations employ participants from WEIRD (western, educated, industrialized, rich, democratic) societies, while evaluating explanations in application domains targeting a diverse range of user types [11]. Prior work has shown that users with different cultural backgrounds have different preferences when it comes to interface design [38], which suggests that the preferences, requirements, and needs might also differ for the format and communication elements of RS explanations.

Collaborations with researchers who investigate the information needs and preferences across different personalities and cultural backgrounds would be a great starting point to further investigate the impact of user and situational characteristics that need to be taken into consideration when evaluating RS explanations.

4. Conclusions

In this paper, we argue that evaluating RS explanations requires a holistic approach that integrates algorithmic and user-centered aspects to cover all four stages of explainability. To address this need, we explored an extension of a theoretical framework to cover all elements. Specifically, we discussed the challenges that we are facing while operationalizing this framework for the evaluation of AI explanations for large-scale evaluation and comparison of RS explanation methods. The main issues that we are facing are related to (1) adapting metrics from a classification problem, (2) finding measures that allow a comparison of intrinsically explainable RS with post-hoc explanation methods at explanation generation and abstraction, and (3) limiting the dimensions of user evaluation at the communication level. Nonetheless, we see the adoption of such a framework into the evaluation practices as a promising way to thoroughly evaluate and compare RS explanations, allowing for systematic investigation of open research gaps, such as the impact of situational and personal characteristics on explanation properties. Future work should look into whether the adoption of a multi-faceted evaluation framework for RS, such as FEVR [39], which covers the evaluation objectives and design space, can help limit the evaluation dimensions while preserving comprehensiveness. We further plan to release a first version of the framework, which extends the Cornac framework [40, 41, 42] with explanation methods and evaluation metrics. Ultimately, we believe that the use of a holistic view on RS explanation evaluations will lead to a better understanding of the impact of various RS explanations, helping both practitioners to build better systems and scientists to be more systematic when exploring novel approaches.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: Grammar and spelling check. After using this tool, the author reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. Raza, C. Ding, News recommender system: a review of recent progress, challenges, and opportunities, Artificial Intelligence Review 55 (2022) 749–800. URL: https://link.springer.com/10.1007/s10462-021-10043-x. doi:10.1007/s10462-021-10043-x.

[2] L. Heitz, J. A. Lischka, R. Abdullah, L. Laugwitz, H. Meyer, A. Bernstein, Deliberative diversity for news recommendations: Operationalization and experimental user study, in: Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 813–819. URL: https://doi.org/10.1145/3604915.3608834. doi:10.1145/3604915.3608834.

[3] M. Bhattacharya, S. Lamkhede, Augmenting netflix search with in-session adapted recommendations, in: Proceedings of the 16th ACM Conference on Recommender Systems, RecSys '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 542–545. URL: https://doi.org/10.1145/3523227.3547407. doi:10.1145/3523227.3547407.

[4] T. Bontempelli, B. Chapus, F. Rigaud, M. Morlon, M. Lorant, G. Salha-Galvan, Flow moods: Recommending music by moods on deezer, in: Proceedings of the 16th ACM Conference on Recommender Systems, RecSys '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 452–455. URL: https://doi.org/10.1145/3523227.3547378. doi:10.1145/3523227.3547378.

[5] Z. Wang, Y. Zou, A. Dai, L. Hou, N. Qiao, L. Zou, M. Ma, Z. Ding, S. Xu, An industrial framework for personalized serendipitous recommendation in e-commerce, in: Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 1015–1018. URL: https://doi.org/10.1145/3604915.3610234. doi:10.1145/3604915.3610234.

[6] J. Dhameliya, N. Desai, Job recommender systems: A survey, in: 2019 Innovations in Power and Advanced Computing Technologies (i-PACT), volume 1, IEEE, Piscataway, New Jersey,USA, 2019, pp. 1–5. doi:10.1109/i-PACT44901.2019.8960231.

[7] N. Tintarev, J. Masthoff, A Survey of Explanations in Recommender Systems, in: 2007 IEEE 23rd International Conference on Data Engineering Workshop, IEEE, Istanbul, Turkey, 2007, pp. 801–810. URL: http://ieeexplore.ieee.org/document/4401070/. doi:10.1109/ICDEW.2007.4401070.

[8] I. Nunes, D. Jannach, A systematic review and taxonomy of explanations in decision support and recommender systems, User Modeling and User-Adapted Interaction 27 (2017) 393–444.

[9] Y. Zhang, X. Chen, et al., Explainable recommendation: A survey and new perspectives, Foundations and Trends® in Information Retrieval 14 (2020) 1–101.

[10] V. Arya, R. K. E. Bellamy, P. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilovic, S. Mourad, P. Pedemonte, R. Raghavendra, J. T. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, Y. Zhang, One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques, 2019. URL: http://arxiv.org/abs/1909.03012. arXiv:1909.03012.

[11] K. Wardatzky, O. Inel, L. Rossetto, A. Bernstein, Whom do explanations serve? a systematic literature survey of user characteristics in explainable recommender systems evaluation, ACM Trans. Recomm. Syst. 3 (2025). URL: https://doi.org/10.1145/3716394. doi:10.1145/3716394.

[12] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38. URL: https://www.sciencedirect.com/science/article/pii/S0004370218305988. doi:https://doi.org/10.1016/j.artint.2018.07.007.

[13] I. Donoso-Guzmán, J. Ooge, D. Parra, K. Verbert, Towards a comprehensive human-centred

evaluation framework for explainable ai, in: L. Longo (Ed.), Explainable Artificial Intelligence, Springer Nature Switzerland, Cham, 2023, pp. 183–204.
[14] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, C. Newell, Explaining the user experience of recommender systems, User Modeling and User-Adapted Interaction 22 (2012) 441–504. URL: https://doi.org/10.1007/s11257-011-9118-4. doi:10.1007/s11257-011-9118-4.
[15] A. Hedström, L. Weber, D. Krakowczyk, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, M. M. M. Höhne, Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond, Journal of Machine Learning Research 24 (2023) 1–11. URL: http://jmlr.org/papers/v24/22-0142.html.
[16] C. Agarwal, S. Krishna, E. Saxena, M. Pawelczyk, N. Johnson, I. Puri, M. Zitnik, H. Lakkaraju, OpenXAI: Towards a transparent evaluation of model explanations, in: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2022. URL: https://openreview.net/forum?id=MU2495w47rz.
[17] L. Coba, R. Confalonieri, M. Zanker, Recoxplainer: A library for development and offline evaluation of explainable recommender systems, IEEE Computational Intelligence Magazine 17 (2022) 46–58. doi:10.1109/MCI.2021.3129958.
[18] A. Ariza-Casabona, L. Boratto, M. Salamó, A comparative analysis of text-based explainable recommender systems, in: Proceedings of the 18th ACM Conference on Recommender Systems, RecSys '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 105–115. URL: https://doi.org/10.1145/3640457.3688069. doi:10.1145/3640457.3688069.
[19] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, 2017. URL: https://arxiv.org/abs/1702.08608. arXiv:1702.08608.
[20] D. V. Carvalho, E. M. Pereira, J. S. Cardoso, Machine learning interpretability: A survey on methods and metrics, Electronics 8 (2019). URL: https://www.mdpi.com/2079-9292/8/8/832. doi:10.3390/electronics8080832.
[21] G. Vilone, L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, Information Fusion 76 (2021) 89–106. URL: https://www.sciencedirect.com/science/article/pii/S1566253521001093. doi:https://doi.org/10.1016/j.inffus.2021.05.009.
[22] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, S. Rinzivillo, Benchmarking and survey of explanation methods for black box models, Data Mining and Knowledge Discovery 37 (2023) 1719–1778. URL: https://doi.org/10.1007/s10618-023-00933-9. doi:10.1007/s10618-023-00933-9.
[23] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai, ACM Comput. Surv. 55 (2023). URL: https://doi.org/10.1145/3583558. doi:10.1145/3583558.
[24] A. R. Mohammadi, A. Peintner, M. Müller, E. Zangerle, Are we explaining the same recommenders? incorporating recommender performance for evaluating explainers, in: Proceedings of the 18th ACM Conference on Recommender Systems, RecSys '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 1113–1118. URL: https://doi.org/10.1145/3640457.3691709. doi:10.1145/3640457.3691709.
[25] L. Coba, P. Symeonidis, M. Zanker, Personalised novel and explainable matrix factorisation, Data & Knowledge Engineering 122 (2019) 142–158. URL: https://www.sciencedirect.com/science/article/pii/S0169023X1830332X. doi:https://doi.org/10.1016/j.datak.2019.06.003.
[26] M. A. Chatti, M. Guesmi, A. Muslim, Visualization for recommendation explainability: A survey and new perspectives, ACM Trans. Interact. Intell. Syst. 14 (2024). URL: https://doi.org/10.1145/3672276. doi:10.1145/3672276.
[27] B. Paudel, F. Christoffel, C. Newell, A. Bernstein, Updatable, accurate, diverse, and scalable recommendations for interactive applications, ACM Trans. Interact. Intell. Syst. 7 (2016). URL: https://doi.org/10.1145/2955101. doi:10.1145/2955101.
[28] G. Balloccu, L. Boratto, G. Fenu, M. Marras, Post processing recommender systems with knowledge graphs for recency, popularity, and diversity of explanations, in: Proceedings of the 45th

International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 646–656. URL: https://doi.org/10.1145/3477495.3532041. doi:10.1145/3477495.3532041.

[29] G. Balloccu, L. Boratto, G. Fenu, M. Marras, Reinforcement recommendation reasoning through knowledge graphs for explanation path quality, Knowledge-Based Systems 260 (2023) 110098. URL: https://www.sciencedirect.com/science/article/pii/S0950705122011947. doi:https://doi.org/10.1016/j.knosys.2022.110098.

[30] L. Li, Y. Zhang, L. Chen, Generate neural template explanations for recommendation, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 755–764. URL: https://doi.org/10.1145/3340531.3411992. doi:10.1145/3340531.3411992.

[31] S. Vrijenhoek, S. Daniil, J. Sandel, L. Hollink, Diversity of what? on the different conceptualizations of diversity in recommender systems, in: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 573–584. URL: https://doi.org/10.1145/3630106.3658926. doi:10.1145/3630106.3658926.

[32] N. Hu, P. A. Pavlou, J. Zhang, On self-selection biases in online product reviews, MIS Quarterly 41 (2017) pp. 449–475. URL: https://www.jstor.org/stable/26629722.

[33] C. K. Manner, W. C. Lane, Who posts online customer reviews? the role of sociodemographics and personality traits, Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior 30 (2017) 19–42. URL: https://www.jcsdcb.com/index.php/JCSDCB/article/view/226.

[34] M. Guesmi, M. A. Chatti, L. Vorgerd, T. Ngo, S. Joarder, Q. U. Ain, A. Muslim, Explaining user models with different levels of detail for transparent recommendation: A user study, in: Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '22 Adjunct, ACM, New York, NY, USA, 2022, p. 175–183. URL: https://doi.org/10.1145%2F3511047.3537685. doi:10.1145/3511047.3537685.

[35] D. Wilkinson, Öznur Alkan, Q. V. Liao, M. Mattetti, I. Vejsbjerg, B. P. Knijnenburg, E. Daly, Why or why not? the effect of justification styles on chatbot recommendations, ACM Transactions on Information Systems 39 (2021) 1–21. URL: https://doi.org/10.1145%2F3441715. doi:10.1145/3441715.

[36] D. C. Hernandez-Bocanegra, J. Ziegler, Explaining review-based recommendations: Effects of profile transparency, presentation style and user characteristics, i-com 19 (2020) 181–200. URL: https://doi.org/10.1515%2Ficom-2020-0021. doi:10.1515/icom-2020-0021.

[37] S. Berkovsky, R. Taib, Y. Hijikata, P. Braslavsku, B. Knijnenburg, A cross-cultural analysis of trust in recommender systems, in: Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization, UMAP '18, ACM, New York, NY, USA, 2018, p. 285–289. URL: https://doi.org/10.1145%2F3209219.3209251. doi:10.1145/3209219.3209251.

[38] K. Reinecke, A. Bernstein, Improving performance, perceived usability, and aesthetics with culturally adaptive user interfaces, ACM Transactions on Computer-Human Interaction 18 (2011) 1–29. URL: https://dl.acm.org/doi/10.1145/1970378.1970382. doi:10.1145/1970378.1970382.

[39] E. Zangerle, C. Bauer, Evaluating recommender systems: Survey and framework, ACM Comput. Surv. 55 (2022). URL: https://doi.org/10.1145/3556536. doi:10.1145/3556536.

[40] A. Salah, Q.-T. Truong, H. W. Lauw, Cornac: A comparative framework for multimodal recommender systems, Journal of Machine Learning Research 21 (2020) 1–5.

[41] Q.-T. Truong, A. Salah, T.-B. Tran, J. Guo, H. W. Lauw, Exploring cross-modality utilization in recommender systems, IEEE Internet Computing (2021).

[42] Q.-T. Truong, A. Salah, H. Lauw, Multi-modal recommender systems: Hands-on exploration, in: Fifteenth ACM Conference on Recommender Systems, 2021, pp. 834–837.