Evaluating the Pros and Cons of Recommender Systems Explanations

Wardatzky, Kathrin

DOI: https://doi.org/10.1145/3640457.3688011

Evaluating the Pros and Cons of Recommender Systems Explanations

Kathrin Wardatzky
wardatzky@ifi.uzh.ch
University of Zurich
Zurich, Switzerland

ABSTRACT

Despite the growing interest in explainable AI in the RecSys community, the evaluation of explanations is still an open research topic. Typically, explanations are evaluated using offline metrics, with a case study, or through a user study. In my research, I will have a closer look at the evaluation of the effects of explanations on users. I investigate two possible factors that can impact the effects reported in recent publications, namely the explanation design and content as well as the users themselves. I further address the problem of determining promising explanations for an application scenario from a seemingly endless pool of options. Lastly, I propose a user study to close some of the research gaps established in the surveys and investigate how recommender systems explanations impact the understanding of users with different backgrounds.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; **Personalization**; • **Human-centered computing**;

KEYWORDS

Explainable AI; Recommender Systems; Evaluation

ACM Reference Format:
Kathrin Wardatzky. 2024. Evaluating the Pros and Cons of Recommender Systems Explanations. In *18th ACM Conference on Recommender Systems (RecSys '24), October 14–18, 2024, Bari, Italy*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3640457.3688011

1 INTRODUCTION

There is currently no commonly agreed-upon definition for AI explanations, and the term explanation is often used interchangeably with explainability or interpretability which is why it is important to clarify the used terminology. I employ the following definition of explanation, adapted from Zhang and Chen [31], for my work: *an explanation is a piece of information that makes it understandable to a user why they are receiving a certain recommendation or how the recommendations were generated.*

Adding explanations to recommender systems can have multiple effects on a user. By increasing the users' understanding of the system, explanations can increase their trust, effectiveness, efficiency, and satisfaction and it has been used as a means to increase the transparency of a recommender system. Additionally, explanations can enable users to correct the system's output if it is wrong and thus provide valuable feedback to the recommender model [24, 27]. However, not all explanations are suitable to evoke these benefits. Recommender systems target a broad and diverse user group with different backgrounds and characteristics who might have different requirements for receiving explanations [20]. Furthermore, the domains in which recommender systems are utilized can vary in the risk that is involved for users if they blindly follow the recommendations which might impact the users' needs to understand the recommendations. These are not the only difficulties that need to be considered when designing explanations. As Balog and Radlinski [6] show, optimizing explanations for a particular effect on a user is challenging, as the effects can correlate. Furthermore, not only the design of explanations with a particular effect in mind is challenging, but also the evaluation of whether this effect was achieved. One particular challenge is the fact that there are currently no standard approaches or guidelines for evaluating explanations [18, 31], which leads to differences in evaluation approaches overall and in the evaluation methodology of one explanation effect. These differences further increase the difficulty of comparing the evaluation results and assessing if they suit another application scenario. These challenges contribute to an unknown state-of-the-art of explanation effects and make it impossible to reproduce reported results. The overarching goal of my PhD project is to address the outlined challenges by *moving toward a comparative evaluation approach for explainable recommender systems* that allows us to determine the *strengths and weaknesses of explanation approaches* as well as understand their *tradeoffs* and the effects they have on different user groups. I target this goal with three work packages. First, I investigate what effects are measured when evaluating recommender systems explanations and what is known about the impact of different conditions on the result. The output of the first work package consists of two systematic literature surveys, in each of which I investigate one factor that might influence the effect of explanations on a user, namely, the users themselves and the design and content of the explanations.

In the second work package, I propose *an end-to-end pipeline covering the entire process of generating end-user-friendly explanations for recommender systems* starting with the selection of the datasets and recommender model, over the explanation approach to the final explanation. This contribution aims to provide a tool to select suitable explanation approaches and explanations to be compared with each other in an evaluation process. The selection of explanation approaches will be determined by multiple evaluation processes at

RecSys '24, October 14–18, 2024, Bari, Italy

ACM ISBN 979-8-4007-0505-2/24/10
https://doi.org/10.1145/3640457.3688011

different steps of the pipeline in order to determine their strengths and weaknesses and narrow down the space of possible options.

In the final work package, I will *extend this pipeline with objective measures of explanation effects on users before evaluating and comparing a set of explanation approaches with a user study.*

2 BACKGROUND

There is a rich body of literature covering different aspects of explainable recommender systems. In this section, I mention relevant work covering the topics of evaluation of AI explanations, the generation and selection of recommender systems explanations for evaluation, and the effects of the explanations.

2.1 Evaluation of AI explanations

Doshi-Velez and Kim [11] distinguish between three evaluation approaches for Machine Learning interpretability which they define as *the ability to explain or present in understandable terms to a human*: functionally-grounded, human-grounded, and application-grounded evaluation. Functionally-grounded evaluation uses a proxy for explanation quality that allows for evaluation that does not require human experiments. In practice, offline metrics can be applied for this type of evaluation. Existing metrics include, but are not limited to, measurements of how well post-hoc explanation methods simulate the recommendation model (e.g., faithfulness [4], and monotonicity [14]), information retrieval-based metrics (e.g., mean explainability precision, and mean explainability recall [1]), or natural language generation metrics which are used to determine how close text-based explanations get to a ground truth dataset (e.g., BLEU and ROUGE as used in [28]). These metrics are often restricted to a specific data type (e.g., textual user reviews as ground truth for BLEU and ROUGE) or the output structure of the explanation method (e.g., input features ranked by their estimated relevance for the black box recommender system output). Apart from the natural language generation metrics, the offline evaluation commonly focuses on the explanation method's performance and not the explanation's quality.

Human-grounded evaluation simplifies an otherwise complex application scenario to the extent that lay humans can perform tasks and assess the explanations. Application-grounded evaluation has domain experts or the actual users of the application evaluate the explanations within the real application. In the context of recommender systems, the actual users of an application are frequently lay humans, which is why there might be an overlap between human- and application-grounded evaluation, and in the following, I will summarize both categories to *user-based evaluation.*

Nauta et al. [16] recently surveyed the evaluation approaches in explainable AI literature and proposed a multi-faceted evaluation approach for explanations based on their findings. The authors argue that interpretability and explainability are multi-faceted concepts and define twelve properties, evaluating explanations on a content, presentation, and user dimension with the goal of moving toward an objective and quantifiable evaluation approach. For my thesis, I will employ the multi-faceted evaluation approach for recommender systems explanations instead of a one-dimensional evaluation.

2.2 Generating and selecting recommender systems explanations for evaluation

Several explainable AI toolkits and frameworks have been published with the goal of generating and evaluating the explanations. To the best of my knowledge, recoXplainer [10] is the first and so far only framework focusing on explainable recommender systems. Their framework offers a selection of four recommendation algorithms combined with three model-specific explanation methods and three model-agnostic post-hoc approaches. All output of the explanation methods is limited to local explanation in either user- or item-style format. Additionally, recoXplainer provides three offline metrics to evaluate the explanations. Mean Explainability Precision (implemented according to [1]) measures the ratio of how many of the items in a top-n recommendation list of a given user are explainable. Model Fidelity is specific for post-hoc explanation approaches using a surrogate explanation model. RecoXplainer provides the implementation according to [19], which measures how well the surrogate model imitates the behavior of the black box recommendation model. As the final metric in this framework, the authors present an Explainability Score, which measures the extent to which latent factors can explain black box recommendations.

Other recommender systems frameworks, such as Cornac [21, 25, 26] or RecBole [30, 32, 33], often contain intrinsically explainable recommender systems, but lack methods to extract the explanations from the algorithms.

General explainable AI frameworks and toolkits, such as IBM's AI Explainability 360 toolkit [5], Captum[1], OpenXAI [2], or InterpretML [17], are usually not specific to an application area and rarely include recommender systems. They generally provide the explanation methods along with a set of metrics to evaluate their output. In contrast to this, XAITK[2] does not contain implemented explanation approaches and metrics. Instead, they provide resources such as publications, existing software, data, resources, and checklists for evaluating explanations. These checklists can be used to evaluate the effects of explanations on the user, which is something that is not covered by the previously mentioned frameworks.

Aside from XAITK, all reviewed frameworks and toolkits focus on evaluating the output of the explanation method, i.e. a list of features that impact the recommendation or the nearest neighbors that have rated the recommended item. This output can be modified in endless ways before being presented to a user. To the best of my knowledge, [8] are the first to address the problem of reducing the design space of explanations for evaluation. They propose an agent that takes explanations as input and predicts human subjects' responses in a specific use case to identify promising or helpful explanations in any given explainable AI use case, which can then be evaluated with a user study. Their approach selects explanations that are likely to help the user with a certain task (forward simulation, model debugging, and counterfactual reasoning), while my work will focus on limiting the explanations to ones that are most likely to evoke a certain effect.

[1] https://captum.ai/
[2] https://xaitk.org/

2.3 Effects of recommender systems explanations

User evaluation of recommender systems explanations usually measures whether they have a certain effect on the user. Tintarev and Masthoff [24] defined seven goals of recommender systems explanations: transparency, scrutability, trust, effectiveness, persuasiveness, efficiency, and satisfaction, which are commonly used as measures in user-based evaluation. These effects can be impacted by the context in which they are evaluated.

One context variable is the users. A few publications evaluate the effect of their explanation approach on users with different characteristics. Chatti et al. [7] evaluate how personal characteristics such as the need for cognition, visualization familiarity, domain knowledge, and technical expertise related to the level of detail the participants prefer for the explanation in a document recommendation scenario. They found that these characteristics impacted the preferred level of detail and concluded that the context in which the users see the recommendation plays a role in determining the level of detail the explanation. Ma et al. [15] found that gender and learning goals impact the effects of explanations in course recommendations for university students. However, the impact of gender on the explanation effect could not be confirmed by Wilkinson et al. [29] in their study on justification styles for chatbot recommendations. Instead, the authors found that age influenced the perceived justification quality. The impact of the Big Five personality traits on the persuasiveness of recommendations using explanations designed according to Cialdini's six persuasive principles (reciprocity, scarcity, authority, social proof, liking, and commitment) [9] has been evaluated by Alslaity and Tran [3], Sofia et al. [23], and Fatahi et al. [12]. Alslaity and Tran [3] found differences between the two evaluated domains—e-commerce and movie recommendations—within the same personality trait group for the majority of the persuasion profiles. Sofia et al. [23] designed justifications following Cialdini's persuasive principles for a music recommender system to promote new artists' songs. They not only found differences in the reception of the justifications between participants with different personalities but also that the participants were not very good at assessing which justification type would be the most persuasive for them. Fatahi et al. [12] aimed to use the explanations to persuade users of movies that they were initially unmotivated to watch. They show that explanations containing influence strategies that are tailored to the respective user personality can successfully persuade them to interact with items that were previously not of interest and plan to extend the analysis to other personality traits, such as the need for cognition. These results show that user characteristics are a factor that needs to be taken into consideration when evaluating and interpreting the effects of explanations.

In terms of user needs for recommender systems explanations, Shang et al. [22] investigate general user needs for counterfactual explanations in everyday recommendations with a survey and an interview study. Their questionnaire found that most participants are interested in receiving explanations for specific everyday recommender applications (e-commerce, point-of-interest, social media, and multimedia). Their following interview study revealed a general need for more explanations and that existing explanations could be too vague or general and lack relevant details. They also found that the cost involved for the user in the decision-making process impacts their explanation needs. However, the application domains they compared are relatively similar in terms of cost and risk for following recommendations.

Kleinerman et al. [13] evaluated their explanations within the same domain (online dating) but experimented with adding costs and incentives to interact with the recommended items. While they could not find a clear preference in the perception of the evaluated recommender systems explanations in the study without added incentives and costs, the follow-up study including these aspects found a clear preference for one explanation type.

Overall, these context variables that can impact the effects of recommender systems explanations are only rarely investigated. During my thesis, I plan to address this gap and further investigate the impact of user characteristics on the measured explanation effects.

3 PROBLEM STATEMENT

In order to move towards a more standardized and comparative evaluation process, I plan to address three main research questions during my PhD.

The first research question is related to the approaches used to evaluate explanations. It is currently unclear *what is known about the factors that impact the measured effects of recommender systems explanations*. I plan to address this in my first research question.

RQ 1: What factors impact the measured effects of a recommendation explanation on a user? This research question addresses the problem of not knowing to what extent factors such as the user characteristics or the design or content of the explanation impact the measured effect. Normally, a meta-study would be the appropriate instrument to assess this research question. This would require a standardized evaluation methodology across publications which is not the case in current research. Instead, I aim to answer this question by systematically analyzing the measured explanation effects in recommender systems and the condition in which they were evaluated with two systematic literature surveys.

RQ 2: How can the design space of possible explanation approaches be reduced to those likely to meet the desired goal? With the second research question, I aim to target the challenge that there are *seemingly endless options* for how an explanation can look like. A practitioner who wants to add explanations to the output of a recommender system might only have a specific budget to try out and test different explanation types. I plan to address this problem by proposing a method to narrow down the possible explanations. It will additionally serve as a tool for the experiment in the final research question.

RQ 3: How do varying conditions impact explanations' effect on a user's understanding? This research question addresses two challenges related to measuring the effects of explanations. The first part of this contribution targets the issue that *explanation effects are often measured by self-assessed perceptions of the effect instead of objective measures*. I plan to extend the method developed in RQ 2 with objective evaluation methods to measure the actual effect of explanations on a user. I will focus this effort on the effect of explanations on a user's understanding of the system to keep

the scope of this contribution to a reasonable level. Furthermore, I need to address the challenge that little is known about *how contextual variables, such as the application domain, impact the effect of explanations*. Therefore, I propose a series of experiments to (1) evaluate how contextual variables, such as the application domain, impact the effect of explanations and (2) collect missing data to test and refine the methodology created to answer the second research question.

4 EVALUATION PLAN AND CONTRIBUTIONS

This section summarizes the evaluation plan and the contributions that I expect to make by answering each research question.

4.1 RQ 1: Investigating what is known

In order to answer the first research question, I conducted two systematic literature reviews of 124 peer-reviewed publications about explainable or interpretable recommender systems. The papers were queried with a search term consisting of synonyms of *explain** combined with *recommend**, selected based on pre-defined inclusion and exclusion criteria, and categorized by their evaluation approach for the explanations. Both surveys focus on explanations that were evaluated by a user. We extracted the reported results from each publication and identified the dependent variable (i.e., the measured effect of the explanation) and the independent variables along with the information that was provided about the explanations and user study design and setup. The first survey paper, currently under review at ACM TORS, analyzes the participants of the user evaluation and how different characteristics of the participants might impact the measured effects. Along with the information about the participants, we extracted the reported results in which either the evaluated effect was measured on different forms of a user characteristic (i.e., male compared to female participants) or a correlation of a characteristic with an effect was measured. We then aggregate the results across publications.

For the second survey, we extracted the reported results in which either a recommender system with an explanation was compared to a system without an explanation or two explanations were compared with each other along with the provided information on the explanations. We categorized the explanations based on how they are displayed, the part of the recommender system they are explaining (i.e., input data, recommendation process, or output), their interactivity, and the type of explanation (i.e., local, global, counterfactual). We aggregated the results separately for results where an explanation was compared to a system without explanation and where multiple explanations were compared with each other to analyze the results.

The surveys conclude with a set of recommendations to mitigate the gaps that were identified in terms of the evaluation methodology of recommender systems explanations.

4.2 RQ 2: Narrowing down the options

The overarching goal of answering this research question is to provide explainable AI researchers with decision support when comparing and evaluating explanation approaches for recommender systems. Two main steps are required to achieve this goal: (1) the development of an end-to-end explanation pipeline for recommender systems explanations that is capable of generating different explanation formats and types and (2) the implementation of multiple evaluation points along the pipeline that allows to rank the explainable recommendation methods based on different criteria.

The generation of the pipeline requires multiple steps. Not every recommender system works with every dataset and not every explanation can be generated with every explanation method. These technical dependencies between data type, recommender system model, explanation generation method, and end-user explanation design need to be mapped out and implemented. The implementation of multiple evaluation points along the pipeline can partially rely on existing evaluation metrics for recommender systems and explanation methods. Additionally, I plan to investigate the impact of structural and distributional aspects of the datasets on the output of explanation methods. I further need to find options to evaluate the quality of the generated explanations. The finished pipeline will be capable of evaluating the strengths and weaknesses of explanation approaches on multiple aspects, which then allows a researcher to make an informed decision on what approaches to experiment with.

4.3 RQ 3: Closing the gaps

The work required to answer the third research question will contribute to closing some of the gaps that were identified by the literature surveys described in Section 4.1. The first part of this contribution is to investigate objective evaluation methods for explanation effects in recommender systems and extend the framework described in Section 4.2 with these methods. I will refer to the existing literature on objectively measuring users' understanding and analyze their advantages and disadvantages before pre-testing and extending the framework with the evaluation measures.

The extended framework will then be used to study recommender systems explanations in varying conditions. The goals are to (1) verify and refine the initial context variables of the framework that might impact explanation effects and (2) move towards a more comparative evaluation of explanations. In order to achieve the first goal, the implemented framework from Section 4.2 will generate a set of different end-user-friendly explanations for the same recommendation and explanation method with data from varying application domains. To avoid learning effects, the user evaluation will follow a between-subjects design. A participant is then confronted with the description of an application domain. The first task is to collect information on their experience and a general understanding of recommender systems in this domain as a baseline before presenting the recommendations and explanations. After seeing the explanations, the participant's understanding is measured again, along with their satisfaction with the explanation. The results of the estimated effect on a user's understanding are then evaluated with the measured effects during the user study. Depending on the results, the model will be refined and re-evaluated.

The second part of this contribution will be a benchmarking experiment comparing multiple explanation approaches using the new estimated effect on understanding along with existing offline metrics.

5 PRELIMINARY RESULTS AND FUTURE WORK

Surveying 124 articles published between 2017 and 2022 revealed multiple challenges.

Analyzing the effects of explanations and how they were measured showed that there is little agreement on how to measure an effect. We extracted the questions or task design with which the effects were measured and found that different aspects were measured under the same term and developed a categorization that accommodates these differences. We additionally found that the majority of evaluations measure an effect by asking the participants to report their perception of whether an explanation evoked a certain effect. This is valid for certain effects that are inherently about the users' opinion, such as satisfaction, or if this is the intended notion of the effect that the researchers want to measure. It should be distinguished between, e.g., objectively measuring whether an explanation contributes to the understanding of a user why they are getting a specific recommendation and asking them to assess this themselves, though.

In terms of the participants of the evaluation, the survey reveals inconsistencies in terms of how the information about the participants is reported which makes it hard to draw conclusions about the external validity of the results and impossible to reproduce the results. It further shows that the majority of the participants sampled in the surveyed studies do not necessarily reflect the user group that is targeted by the recommender system. We found predominantly participants from so-called WEIRD (white, educated, industrialized, rich, democratic) societies, which needs to be taken into consideration when analyzing what we know about the effects that recommender systems explanations have on users.

The impact of different user characteristics on the measured effects is rarely investigated. There are, for example, no evaluations of whether a user's demographic information, such as age or gender, impacts the measured effectiveness and efficiency that an explanation contributes to the recommendations. We did find some indications for user characteristics that might have an impact on the explanation effects. Every time the data was disaggregated by the participants' level of social awareness, an impact on trust and transparency was found.

When analyzing the modalities of the recommender systems explanations, we found one explanation type to be particularly dominant: local explanations displayed as text that focuses on explaining the output of the recommender system. We analyzed a total number of 458 explanations, out of which 46% of the explanations are of this type. This means that the majority of the information that we know about explanation effects might be limited to this explanation type, and further research is required to investigate the effect that other explanation types might have on a user.

These results impose some challenges but also opportunities for my future work. The imbalanced data in terms of the effect of explanation modalities and the little knowledge of the impact of user characteristics on the measured effects of explanations need to be accommodated in both work packages. At the same time, the identified gaps and challenges provide opportunities for further experimentation that can have a meaningful impact on the state-of-the-art.

6 CONCLUSIONS

In this paper, I outline a research plan to move toward a comparative evaluation approach for explainable recommender systems that allows us to determine the strengths and weaknesses of the explanation approaches as well as understand their tradeoffs and the effects they have on different user groups. I proposed three contributions in which I investigate (1) what is known about the effects of recommender systems explanations and their interactions with user characteristics, (2) a method to generate explanations that are likely to evoke a certain effect, and (3) a user study to close some of the gaps established in the first contribution.

ACKNOWLEDGMENTS

I would like to thank my advisors, Prof. Abraham Bernstein, PhD, Dr. Oana Inel, and Dr. Luca Rossetto, for their continuous support.

REFERENCES

[1] Behnoush Abdollahi and Olfa Nasraoui. 2017. Using Explainability for Constrained Matrix Factorization. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, Como Italy, 79–83. https://doi.org/10.1145/3109859.3109913

[2] Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. 2022. OpenXAI: Towards a Transparent Evaluation of Model Explanations. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=MU2495w47rz

[3] Alaa Alslaity and Thomas Tran. 2020. The effect of personality traits on persuading recommender system users. In *Conference on recommender systems*.

[4] David Alvarez Melis and Tommi Jaakkola. 2018. Towards Robust Interpretability with Self-Explaining Neural Networks. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf

[5] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2019. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. https://arxiv.org/abs/1909.03012

[6] Krisztian Balog and Filip Radlinski. 2020. Measuring Recommendation Explanation Quality: The Conflicting Goals of Explanations. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Virtual Event China, 329–338. https://doi.org/10.1145/3397271.3401032

[7] Mohamed Amine Chatti, Mouadh Guesmi, Laura Vorgerd, Thao Ngo, Shoeb Joarder, Qurat Ul Ain, and Arham Muslim. 2022. Is More Always Better? The Effects of Personal Characteristics and Level of Detail on the Perception of Explanations in a Recommender System. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, Barcelona Spain, 254–264. https://doi.org/10.1145/3503252.3531304

[8] Valerie Chen, Nari Johnson, Nicholay Topin, Gregory Plumb, and Ameet Talwalkar. 2022. Use-Case-Grounded Simulations for Explanation Evaluation. (2022). https://doi.org/10.48550/ARXIV.2206.02256 Publisher: arXiv Version Number: 2.

[9] Robert B Cialdini. 2009. *Influence: Science and practice*. Vol. 4.

[10] Ludovik Coba, Roberto Confalonieri, and Markus Zanker. 2022. RecoXplainer: A Library for Development and Offline Evaluation of Explainable Recommender Systems. *IEEE Computational Intelligence Magazine* 17, 1 (Feb. 2022), 46–58. https://doi.org/10.1109/MCI.2021.3129958

[11] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. (2017). https://doi.org/10.48550/ARXIV.1702.08608 Publisher: arXiv Version Number: 2.

[12] Somayeh Fatahi, Mina Mousavifar, and Julita Vassileva. 2023. Investigating the effectiveness of persuasive justification messages in fair music recommender systems for users with different personality traits. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. 66–77.

[13] Akiva Kleinerman, Ariel Rosenfeld, Francesco Ricci, and Sarit Kraus. 2021. Supporting users in finding successful matches in reciprocal recommender systems. *User Modeling and User-Adapted Interaction* 31, 3 (July 2021), 541–589. https://doi.org/10.1007/s11257-020-09279-z

[14] Ronny Luss, Pin-Yu Chen, Amit Dhurandhar, Prasanna Sattigeri, Karthikeyan Shanmugam, and Chun-Chen Tu. 2019. Generating Contrastive Explanations with Monotonic Attribute Functions. *CoRR* abs/1905.12698 (2019). arXiv:1905.12698 http://arxiv.org/abs/1905.12698

[15] Boxuan Ma, Min Lu, Yuta Taniguchi, and Shin'ichi Konomi. 2021. CourseQ: the impact of visual and interactive course recommendation in university environments. *Research and Practice in Technology Enhanced Learning* 16, 1 (Dec. 2021), 18. https://doi.org/10.1186/s41039-021-00167-7

[16] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv.* 55, 13s, Article 295 (jul 2023), 42 pages. https://doi.org/10.1145/3583558

[17] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv preprint arXiv:1909.09223* (2019).

[18] Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction* 27, 3-5 (Dec. 2017), 393–444. https://doi.org/10.1007/s11257-017-9195-0

[19] Georgina Peake and Jun Wang. 2018. Explanation Mining: Post Hoc Interpretability of Latent Factor Models for Recommendation Systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, London United Kingdom, 2060–2069. https://doi.org/10.1145/3219819.3220072

[20] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2015. Recommender Systems: Introduction and Challenges. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, Boston, MA, 1–34. https://doi.org/10.1007/978-1-4899-7637-6_1

[21] Aghiles Salah, Quoc-Tuan Truong, and Hady W Lauw. 2020. Cornac: A Comparative Framework for Multimodal Recommender Systems. *Journal of Machine Learning Research* 21, 95 (2020), 1–5.

[22] Ruoxi Shang, K. J. Kevin Feng, and Chirag Shah. 2022. Why Am I Not Seeing It? Understanding Users' Needs for Counterfactual Explanations in Everyday Recommendations. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1330–1340. https://doi.org/10.1145/3531146.3533189

[23] Gkika Sofia, Skiada Marianna, Lekakos George, and Kourouthanasis Panos. 2016. Investigating the role of personality traits and influence strategies on the persuasive effect of personalized recommendations. In *4th Workshop on emotions and personality in personalized systems (EMPIRE)*, Vol. 9.

[24] Nava Tintarev and Judith Masthoff. 2007. A survey of explanations in recommender systems. In *2007 IEEE 23rd international conference on data engineering workshop*. IEEE, 801–810.

[25] Quoc-Tuan Truong, Aghiles Salah, and Hady Lauw. 2021. Multi-modal recommender systems: Hands-on exploration. In *Fifteenth ACM Conference on Recommender Systems*. 834–837.

[26] Quoc-Tuan Truong, Aghiles Salah, Thanh-Binh Tran, Jingyao Guo, and Hady W Lauw. 2021. Exploring Cross-Modality Utilization in Recommender Systems. *IEEE Internet Computing* (2021).

[27] Alexandra Vultureanu-Albişi and Costin Bădică. 2022. A survey on effects of adding explanations to recommender systems. *Concurrency and Computation: Practice and Experience* (Jan. 2022). https://doi.org/10.1002/cpe.6834

[28] Peng Wang, Renqin Cai, and Hongning Wang. 2022. Graph-Based Extractive Explainer for Recommendations. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 2163–2171. https://doi.org/10.1145/3485447.3512168

[29] Daricia Wilkinson, Öznur Alkan, Q. Vera Liao, Massimiliano Mattetti, Inge Vejsbjerg, Bart P. Knijnenburg, and Elizabeth Daly. 2021. Why or Why Not? The Effect of Justification Styles on Chatbot Recommendations. *ACM Transactions on Information Systems* 39, 4 (Oct. 2021), 1–21. https://doi.org/10.1145/3441715

[30] Lanling Xu, Zhen Tian, Gaowei Zhang, Junjie Zhang, Lei Wang, Bowen Zheng, Yifan Li, Jiakai Tang, Zeyu Zhang, Yupeng Hou, Xingyu Pan, Wayne Xin Zhao, Xu Chen, and Ji-Rong Wen. 2023. Towards a More User-Friendly and Easy-to-Use Benchmark Library for Recommender Systems. In *SIGIR*. ACM, 2837–2847.

[31] Yongfeng Zhang and Xu Chen. 2020. Explainable Recommendation: A Survey and New Perspectives. *Foundations and Trends® in Information Retrieval* 14, 1 (2020), 1–101. https://doi.org/10.1561/1500000066

[32] Wayne Xin Zhao, Yupeng Hou, Xingyu Pan, Chen Yang, Zeyu Zhang, Zihan Lin, Jingsen Zhang, Shuqing Bian, Jiakai Tang, Wenqi Sun, Yushuo Chen, Lanling Xu, Gaowei Zhang, Zhen Tian, Changxin Tian, Shanlei Mu, Xinyan Fan, Xu Chen, and Ji-Rong Wen. 2022. RecBole 2.0: Towards a More Up-to-Date Recommendation Library. In *CIKM*. ACM, 4722–4726.

[33] 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. 2021. RecBole: Towards a Unified, Comprehensive and Efficient Framework for Recommendation Algorithms. In *CIKM*. ACM, 4653–4664.