Towards Transparent Recommender Systems via Argumentation Frameworks

Elena Stefancova

Comenius University Bratislava, Slovakia

Abstract

As artificial intelligence becomes more widespread, ensuring its trustworthiness is increasingly important. This work focuses on enhancing transparency in recommender systems by analyzing key fairness dimensions and their trade-offs, aiming to improve user trust. We also propose a novel synthetic data generation method to study changes in internal representations, offering insights into system behavior and decision-making. Ongoing work explores explainability, potentially via argumentation frameworks, to further support transparent and accountable recommendations.

Keywords

recommender systems, fairness, explainability, argumentation frameworks

1. Introduction

With the growing ubiquity of artificial intelligence, concerns around its trustworthiness grow accordingly. The European Union's Artificial Intelligence Act [1] recognizes these risks by classifying certain types of recommender systems as high-risk AI.

Recommender systems are widely used in online platforms to personalize content and reduce information overload. Given their growing impact, ensuring their trustworthiness is essential. Eliminating these systems is not feasible, making it crucial to address associated ethical challenges such as transparency.

Trustworthiness in AI is the focus of a dedicated research community. Two critical pillars of trustworthiness are explainability and fairness. Explainability enables stakeholders to understand and evaluate system decisions, fostering trust. Fairness ensures equitable treatment of different user groups, preventing bias and supporting ethical AI adoption.

Due to the complex, multi-stakeholder nature of recommendation tasks, fairness often involves trade-offs between competing criteria. Additionally, real-world datasets used for evaluating fairness are often limited or biased. To address this, we explore the fairness-accuracy tradeoffs and introduce a method for generating synthetic data, which allows for controlled experiments on system behavior and fairness interventions.

The main goal of this work is to enhance the trustworthiness of recommender systems. This is achieved by:

- developing a synthetic data generation method for evaluating changes in internal representations;
- analyzing fairness dimensions and their trade-offs;
- improving system interpretability to strengthen user trust.

The outcomes contribute toward more transparent, fair, and explainable recommender systems.

To improve system interpretability, current efforts are exploring explainability through potential integration with argumentation-based frameworks.

Doctoral Consortium of the 22nd International Conference on Principles of Knowledge Representation and Reasoning (KR 2025 DC), November 11-17, 2025, Melbourne, Australia

stefancova27@uniba.sk (E. Stefancova)

0000-0001-8683-939X (E. Stefancova)

2. Related Work

Recent research has shifted from purely accuracy-oriented recommender systems towards more trustworthy models that incorporate principles such as fairness, transparency, explainability, and robustness [2]. A trustworthy recommender system must not only deliver relevant results but also be secure, responsible, and understandable to stakeholders. Transparency and explainability are especially critical, as they improve user confidence and system accountability [3].

The development of trustworthy systems spans multiple stages, including robust and explainable data representation, fair and transparent recommendation generation, and ethically grounded evaluation practices. Challenges such as noisy or biased data remain significant, and synthetic data generation has emerged as a potential solution for controlled experimentation and improving reliability [4].

This growing body of work highlights the importance of designing recommendation pipelines that are not only effective, but also aligned with broader ethical and regulatory frameworks, such as the EU's Artificial Intelligence Act.

2.1. Fairness

Fairness in recommender systems has emerged as a critical aspect of trustworthy AI, aiming to mitigate various forms of bias embedded in the data, models, and feedback loops [5]. These biases can take the form of data (e.g., exposure, cold-start, popularity), model (e.g., ranking), or feedback biases [6]. While fairness-aware techniques can reduce discrimination and improve the treatment of underrepresented groups, they may introduce trade-offs with traditional accuracy metrics [2]. In some cases, fairness efforts also lead to beneficial side effects, such as improved diversity or better coverage of the long tail [7].

The concept of fairness is highly contextual and multi-dimensional [8]. Numerous definitions have been proposed, including group fairness, individual fairness, process fairness, and outcome fairness [9]. However, most research focuses on a single, static definition of fairness, often failing to reflect the complexity of real-world deployments with multiple stakeholder groups. Fairness considerations typically focus more on providers (e.g., item suppliers) than consumers (e.g., users), even though both perspectives are necessary for holistic assessment [9]. While some studies consider personalized fairness based on users' preferences or histories [10], these approaches rarely combine multiple fairness definitions or adapt to dynamic user contexts.

Fairness-aware recommendation methods span the full system pipeline, from pre-processing (e.g., data relabeling or reweighting), in-processing (e.g., fairness constraints during training), to post-processing (e.g., re-ranking) [6]. These methods can be static, aiming for fairness across a population, or dynamic, tailoring fairness to each user interaction [11]. The choice between static and dynamic fairness often depends on the application context—for example, newsletter recommendations benefit from static batch approaches, while real-time interfaces require adaptive strategies. Evaluation metrics also vary, including global fairness, group proportional fairness, mean reciprocal rank fairness, and others, each capturing different trade-offs and fairness goals [8]. Despite recent progress, challenges remain due to the multiplicity of fairness definitions, limited real-world deployments, and insufficient evaluation across diverse stakeholder needs [12].

2.2. Explainability

Explainability plays a crucial role in fostering the user trust, enabling users and developers alike to assess system behavior, identify biases, and ensure fair and ethical use [13].

Explainable AI (XAI) encompasses techniques that either make models inherently interpretable (white-box models) or provide post-hoc explanations for opaque, black-box systems [14]. Techniques such as LIME [15] and SHAP [16] offer local or global insights into decision-making processes. These approaches serve purposes ranging from model debugging to improving user confidence and meeting

regulatory demands [17]. Nevertheless, post-hoc explainers often come with trade-offs in fidelity and may not fully reflect the internal model logic [18].

Within recommender systems, explainability is approached through both interpretable models (e.g., matrix factorization, knowledge-based approaches) and post-hoc methods (e.g., attention mechanisms, rule extraction). A well-established taxonomy includes user-based, item-based, feature-based, and opinion-based explanations [19]. These explanations contribute not only to transparency but also to system persuasiveness, user satisfaction, and effectiveness [20].

These provide users with reasons why one item was recommended over another, supporting more informed decision-making and improving user comprehension of ranking logic. By framing recommendations in relative terms, these methods align system outputs with human reasoning and enhance user interaction [21].

The continued exploration of explainability frameworks remains essential for building recommender systems that are not only effective but also responsible and comprehensible.

Argumentation frameworks have increasingly been recognized as a powerful tool for enhancing explainability in artificial intelligence systems. These frameworks provide a structured way to model and evaluate conflicting information, allowing AI models to generate human-understandable explanations by simulating argumentative dialogues or reasoning processes [22]. In the context of explainable AI, argumentation frameworks serve to clarify decision-making processes by explicitly representing supporting and opposing arguments related to specific predictions or recommendations [23]. This approach not only aids transparency but also facilitates trust by enabling users to engage interactively with the system's reasoning, thereby improving interpretability and user acceptance.

Moreover, argumentation-based explainability has been applied to various domains, including recommender systems, where it supports multi-stakeholder perspectives by incorporating diverse viewpoints and fairness considerations into the explanation generation process [24]. Despite its potential, challenges remain in scaling argumentation frameworks to complex models and integrating them seamlessly with existing AI architectures, highlighting ongoing research efforts to balance computational efficiency with explanatory depth.

Overall, argumentation frameworks represent a promising avenue for advancing explainability by enabling AI systems to provide nuanced, interactive, and context-aware justifications for their outputs.

3. Motivation

My motivation to engage deeply with this research topic stems from a strong alignment between my personal interests and the strategic focus of my research groups. Over the past nine years, I have cultivated extensive experience in recommender systems, beginning with my Master's thesis and subsequently exploring the ethical dimensions of these systems in a professional context. During this period, I concentrated on identifying and mitigating the negative effects of recommender systems, such as filter bubbles and related biases.

During my doctoral studies at Comenius University in Bratislava, Slovakia, I became a member of a research group specializing in knowledge representation and explainable artificial intelligence, which remains my primary academic affiliation. Concurrently, I had the opportunity to spend an academic year at the University of Colorado Boulder, USA, supported by a Fulbright Award. There, I collaborated with a research group focused on recommender systems, particularly investigating fairness considerations from the perspectives of multiple stakeholders. Our work also emphasizes the critical need for model transparency and effective communication regarding the rationale behind recommendations, especially when fairness constraints influence the outcomes.

Currently, my research agenda prioritized advancing fairness-aware recommendation systems while progressively shifting focus towards explainability. Having completed the initial phase centered on fairness, I am dedicated to developing explainability techniques and refining argumentation frameworks, ultimately aiming to complete my doctoral dissertation.

Participating in the doctoral consortium would provide invaluable feedback, particularly on aspects

related to explainability and the integration of argumentation frameworks, thereby enriching the rigor and impact of my research.

4. Project Proposal

My project proposal consists of several steps.

Firstly, fairness integration. Employing a hybrid multi-stage recommendation architecture where an initial personalized recommendation list is re-ranked to incorporate fairness constraints. This approach supports multiple, dynamic fairness definitions via allocation and choice mechanisms, enabling modular, scalable, and partly explainable adjustments. The research investigates trade-offs between fairness and accuracy and the joint effects of various fairness notions, including provider-side, consumer-side, group, and individual fairness.

Secondly, developing a synthetic data generation method inspired by matrix factorization techniques to simulate realistic user-item interactions and controlled biases. This allows systematic evaluation of fairness-aware algorithms under diverse conditions and supports experimentation on fairness-accuracy trade-offs with adjustable sensitive features and bias parameters.

Lastly, enhancing explainability by leveraging knowledge representation methods such as ontologies, argumentation frameworks, and comparative explanations. The work focuses on communicating why certain items are recommended, especially when fairness mechanisms affect rankings. Comparative explanations provide transparency regarding deviations introduced by fairness-aware re-ranking. User studies are planned to evaluate explanation effectiveness, assessing metrics such as trust, transparency, and user satisfaction.

Employing the argumentation frameworks allows the system to articulate not only why certain items were recommended but also how competing fairness considerations and stakeholder preferences were balanced. Moreover, argumentation frameworks facilitate the generation of dynamic, context-aware explanations that can be tailored to different stakeholder groups, thereby improving the communicative clarity and relevance of explanations. Through this explicit reasoning process, users can better comprehend the trade-offs and decisions embedded in the recommendation, which fosters acceptance of the system's outputs.

5. Preliminary Results

We explored the research questions through **SCRUF-D**, a dynamic multi-agent framework for fairness-aware recommendation [25]. Unlike static approaches, SCRUF-D models fairness as a dynamic property, using multiple *fairness agents*, each representing a different fairness objective (e.g., group proportionality, utility). Agents are allocated using mechanisms such as *Lottery*, *Least Fair*, or *Weighted*, and their preferences are aggregated through social choice methods (e.g., *Borda*, *Copeland*, *Rescoring*).

Our experiments on real-world (e.g., MovieLens [26], Microlending [27]) and synthetic datasets show that SCRUF-D can effectively balance multiple, heterogeneous fairness definitions with minimal accuracy trade-offs. Our research has shown that it supports both group and individual provider-side fairness [27].

To enable controlled experimentation, we developed **LAFS**, a synthetic data generation method based on latent factor simulation [28]. LAFS simulates user/item features, biases, and sensitive attributes, making it well-suited for testing fairness interventions in recommendation pipelines.

SCRUF-D demonstrates strong flexibility across allocation and ranking mechanisms, enabling nuanced fairness control across dynamic and multi-objective contexts.

6. Conclusion

This work contributes to the development of transparent and fair recommender systems by addressing multiple aspects of trustworthiness, including fairness-aware re-ranking, dynamic multi-objective

optimization, and synthetic data generation. The SCRUF-D framework enables modular and flexible integration of fairness objectives, while the LAFS method allows for systematic evaluation through controlled synthetic datasets.

While these contributions offer a foundation for fair and accountable recommendations, future work is shifting focus toward the dimension of explainability. In particular, we are exploring argumentation-based frameworks as a structured approach for modeling, communicating, and justifying the reasoning behind recommendations. Argumentation offers a promising pathway for incorporating stakeholder perspectives, resolving conflicts between competing fairness goals, and providing interactive, context-aware explanations. Ultimately, we aim to integrate fairness and explainability into a unified framework that supports transparent, intelligible, and socially responsible recommendation processes.

In particular, the enhancement of explainability through the incorporation of argumentation frameworks would significantly benefit from the opportunity to participate in the KR Doctoral Consortium and to receive expert feedback.

Acknowledgments

Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I05-03-V02-00064.

Declaration on Generative AI

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

References

[1] European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Technical Report, European Commission, 2021. URL: https://artificialintelligenceact.eu.

[2] S. Wang, X. Zhang, Y. Wang, F. Ricci, Trustworthy recommender systems, ACM Trans. Intell. Syst. Technol. 15 (2024). URL: https://doi.org/10.1145/3627826. doi:10.1145/3627826.

[3] R. Sinha, K. Swearingen, The role of transparency in recommender systems, in: CHI '02 Extended Abstracts on Human Factors in Computing Systems, CHI EA '02, Association for Computing Machinery, New York, NY, USA, 2002, p. 830–831. URL: https://doi.org/10.1145/506443.506619. doi:10.1145/506443.506619.

[4] M. Slokom, Comparing recommender systems using synthetic data, in: Proceedings of the 12th ACM Conference on Recommender Systems, RecSys '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 548–552. URL: https://doi.org/10.1145/3240323.3240325. doi:10.1145/3240323.3240325.

[5] D. Pedreshi, S. Ruggieri, F. Turini, Discrimination-aware data mining, in: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2008, pp. 560–568.

[6] Y. Wang, W. Ma, M. Zhang, Y. Liu, S. Ma, A survey on the fairness of recommender systems, ACM Trans. Inf. Syst. 41 (2023). URL: https://doi.org/10.1145/3547333. doi:10.1145/3547333.

[7] N. Ranjbar Kermany, W. Zhao, J. Yang, J. Wu, L. Pizzato, A fairness-aware multi-stakeholder recommender system, World Wide Web 24 (2021) 1995–2018.

[8] J. J. Smith, A. Buhayh, A. Kathait, P. Ragothaman, N. Mattei, R. Burke, A. Voida, The many faces of fairness: Exploring the institutional logics of multistakeholder microlending recommendation, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023, pp. 1652–1663.

[9] M. D. Ekstrand, A. Das, R. Burke, F. Diaz, et al., Fairness in information access systems, Foundations and Trends® in Information Retrieval 16 (2022) 1–177.
[10] W. Liu, R. Burke, Personalizing fairness-aware re-ranking, arXiv preprint arXiv:1809.02921 (2018). Presented at the 2nd FATRec Workshop held at RecSys 2018, Vancouver, CA.
[11] G. K. Patro, A. Biswas, N. Ganguly, K. P. Gummadi, A. Chakraborty, Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms, in: Proceedings of The Web Conference 2020, 2020, pp. 1194–1204.
[12] H. Cramer, K. Holstein, J. W. Vaughan, H. Daumé III, M. Dudík, H. Wallach, S. Reddy, J. Garcia-Gathright, Challenges of incorporating algorithmic fairness into industry practice, 2019.
[13] A. Ferrario, M. Loi, How explainability contributes to trust in ai, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 1457â€“1466. URL: https://doi.org/10.1145/3531146.3533202. doi:10.1145/3531146.3533202.
[14] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature machine intelligence 1 (2019) 206–215.
[15] M. T. Ribeiro, S. Singh, C. Guestrin, "why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, 2016, pp. 1135–1144.
[16] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.
[17] P. Gohel, P. Singh, M. Mohanty, Explainable ai: current status and future directions, arXiv preprint arXiv:2107.07045 (2021).
[18] C. Musto, M. de Gemmis, P. Lops, G. Semeraro, Generating post hoc review-based natural language justifications for recommender systems, User Modeling and User-Adapted Interaction 31 (2021) 629–673. URL: https://doi.org/10.1007/s11257-020-09270-8. doi:10.1007/s11257-020-09270-8.
[19] Y. Zhang, X. Chen, Explainable recommendation: A survey and new perspectives, Found. Trends Inf. Retr. 14 (2020) 1–101. URL: https://doi.org/10.1561/1500000066. doi:10.1561/1500000066.
[20] N. Tintarev, J. Masthoff, A survey of explanations in recommender systems, in: Data Engineering Workshop, 2007 IEEE 23rd International Conference on, IEEE, 2007, pp. 801–810.
[21] A. Yang, N. Wang, R. Cai, H. Deng, H. Wang, Comparative explanations of recommendations, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 3113–3123. doi:10.1145/3485447.3512031.
[22] A. Vassiliades, N. Bassiliades, T. Patkos, Argumentation and explainable artificial intelligence: a survey, The Knowledge Engineering Review 36 (2021) e5. doi:10.1017/S0269888921000011.
[23] T. Kampik, K. Čyras, J. Ruiz Alarcón, Change in quantitative bipolar argumentation: Sufficient, necessary, and counterfactual explanations, International Journal of Approximate Reasoning 164 (2024) 109066. URL: https://www.sciencedirect.com/science/article/pii/S0888613X23001974. doi:https://doi.org/10.1016/j.ijar.2023.109066.
[24] S. Naveed, T. Donkers, J. Ziegler, Argumentation-based explanations in recommender systems: Conceptual framework and empirical results, in: Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, UMAP '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 293–298. URL: https://doi.org/10.1145/3213586.3225240. doi:10.1145/3213586.3225240.
[25] A. Aird, P. Farastu, J. Sun, E. Stefancová, C. All, A. Voida, N. Mattei, R. Burke, Dynamic fairness-aware recommendation through multi-agent social choice, ACM Trans. Recomm. Syst. 3 (2024). URL: https://doi.org/10.1145/3690653. doi:10.1145/3690653.
[26] F. M. Harper, J. A. Konstan, The movielens datasets: History and context, ACM Transactions on Interactive Intelligent Systems (TiiS) 5 (2015) 19.
[27] A. Aird, E. Štefancová, C. All, A. Voida, M. Homola, N. Mattei, R. Burke, Social choice for

heterogeneous fairness in recommendation, in: Proceedings of the 18th ACM Conference on Recommender Systems, RecSys '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 1096–1101. URL: https://doi.org/10.1145/3640457.3691706. doi:10.1145/3640457.3691706.

[28] E. Stefancova, C. All, J. Paup, M. Homola, N. Mattei, R. Burke, Data generation via latent factor simulation for fairness-aware re-ranking, Presented at the 2024 FAccTRec Workshop on Responsible Recommendation, 2024. URL: https://arxiv.org/abs/2409.14078. arXiv:2409.14078.