Building User Trust in Recommendations via Fairness and Explanations

Dimitris Sacharidis
E-Commerce Research Unit
TU Wien
Vienna, Austria
dimitris@ec.tuwien.ac.at

ABSTRACT

Modern *Artificial Intelligence* (AI) techniques, based on the statistical analysis of big volumes of data, are quickly gaining traction across various domains. *Recommender Systems* are a class of AI techniques that extract preference patterns from large traces of human behavior. Recommenders assist people in taking decisions that range from harmless everyday life dilemmas, e.g., what shoes to buy, to seemingly innocuous choices but with long-term, hidden consequences, e.g., what news article to read, up to more critical decisions, e.g., which person to hire.

As more and more aspects of our everyday lives are influenced by automated decisions made by recommender systems, it becomes natural to question whether these systems are *trustworthy*, particularly given the opaqueness and complexity of their internal workings. These questions are timely posed in the broader context of concerns regarding the societal and ethical implications of applying AI techniques, which have also brought about new regulations, like the EU's "Right to Explanation" [2].

In this talk, we discuss techniques for increasing the user's trust in the decisions of a recommender system, focusing on *fairness* aspects and *explanation* approaches. On the one hand, fairness means that the system exhibits certain desirable ethical traits, such as being non-discriminatory, diversity-aware, and bias-free. On the other hand, explanations provide human-understandable interpretations of the inner working of the system. Both mechanisms can be used in tandem to promote trust in the system. In addition, we investigate user trust from the standpoint of different stakeholders that potentially have varying levels of technical background and diverse needs.

The concept of fairness in AI techniques can be operationalized in various forms [3]. As this research field is still in its early stages, it lacks clarity and consistency, with each work introducing a new definition of fairness. This talk attempts to categorize fairness concerns that may lead to low trust in recommender systems. We start with the observation that fairness means a lack of discrimination when allocating some sensitive resource. In the context of recommender systems, we distinguish between two resources: *accuracy* and *presentation*. In the former, as in [8], the concern is about the relevance of the recommendations, or equivalently about how the system *treats* the various stakeholders. In the latter, as in [5], the concern is about the effects of making recommendations, or equivalently about how the system *impacts* the various stakeholders. In recommender systems, stakeholders are the *consumers* of recommendations, the *providers* of the objects of recommendations, and the *owners* of the system [1].

Explanations seek to make AI systems more trustworthy. For recommender systems, the most common purposes of an explanation are: to show how the system works (*transparency*); to help users make good decisions (*effectiveness*); to increase trust in the system; to convince users (*persuasiveness*); to increase users' perceived *satisfaction*; to enable users to tell the system it is wrong (*scrutability*) [6]. Modern recommenders are based on model-based collaborative techniques, such as matrix factorization and deep neural networks, that have a large number of parameters not directly interpretable. Therefore, it is necessary to employ black box techniques to provide plausible interpretations to recommender outputs. One relevant line of research is on proxy models that approximate system decisions with a simpler interpretable model, as in [4]. Another direction is counterfactual explanations that present examples where the opposite decision would be observed [7].

We highlight two shortcomings of existing work on recommendation explanations. First, they consider a single stakeholder, the consumer of recommendations, and are not suitable for building trust of multiple stakeholders with different concerns. Second, they seek to explain a single recommendation at a time, rather that observing and explaining the long-term behavior of system.

CCS CONCEPTS

• **Information systems → Recommender systems**.

KEYWORDS

Recommender Systems; User Trust; Fairness; Explanations

ACM Reference Format:
Dimitris Sacharidis. 2020. Building User Trust in Recommendations via Fairness and Explanations. In *Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct), July 14–17, 2020, Genoa, Italy*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3386392.3399995

UMAP '20 Adjunct, July 14–17, 2020, Genoa, Italy

ACM ISBN 978-1-4503-7950-2/20/07.
https://doi.org/10.1145/3386392.3399995

BIOGRAPHY

Dimitris Sacharidis is with the E-Commerce Research Unit of TU Wien, Austria. Prior to that, he was a junior researcher at "Athena" Research Center, Greece, and a postdoc at the Hong Kong University of Science and Technology supported by a Marie Skłodowska-Curie individual fellowship. His research interests include topics related to data science, data management, data mining, and recommender systems.

REFERENCES

[1] Robin Burke. 2017. Multisided Fairness for Recommendation. *CoRR* abs/1707.00093 (2017).

[2] Bryce Goodman and Seth R. Flaxman. 2017. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine* 38, 3 (2017), 50–57.

[3] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA.*

[4] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *KDD.*

[5] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *KDD.* ACM, 2219–2228. https://doi.org/10.1145/3219819.3220088

[6] Nava Tintarev and Judith Masthoff. 2007. A Survey of Explanations in Recommender Systems. In *ICDEW.*

[7] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *CoRR* abs/1711.00399 (2017).

[8] Sirui Yao and Bert Huang. 2017. Beyond Parity: Fairness Objectives for Collaborative Filtering. In *NIPS.* 2925–2934.