Interactivity, Fairness and Explanations in Recommendations

Giorgos Giannopoulos*
IMSI/Athena Research Center
Athens, Greece
giann@athenarc.gr

George Papastefanatos
IMSI/Athena Research Center
Athens, Greece
gpapas@athenarc.gr

Dimitris Sacharidis
Université Libre de Bruxelles
Brussels, Belgium
dimitris.sacharidis@ulb.be

Kostas Stefanidis
Tampere University
Tampere, Finland
konstantinos.stefanidis@tuni.fi

ABSTRACT

More and more aspects of our everyday lives are influenced by automated decisions made by systems that statistically analyze traces of our activities. It is thus natural to question whether such systems are trustworthy, particularly given the opaqueness and complexity of their internal workings. In this paper, we present our ongoing work towards a framework that aims to increase trust in machine-generated recommendations by combining ideas from three separate recent research directions, namely *explainability*, *fairness* and *user interactive visualization*. The goal is to enable different stakeholders, with potentially varying levels of background and diverse needs, to query, understand, and fix sources of distrust.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Human-centered computing** → **Interaction design process and methods**; **Visualization**.

KEYWORDS

Recommender systems, explainability, fairness, interactive visualization

ACM Reference Format:
Giorgos Giannopoulos, George Papastefanatos, Dimitris Sacharidis, and Kostas Stefanidis. 2021. Interactivity, Fairness and Explanations in Recommendations. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21 Adjunct), June 21–25, 2021, Utrecht, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3450614.3462238

1 INTRODUCTION

Modern *Artificial Intelligence* (AI) techniques, based on the statistical analysis of big volumes of data, are quickly gaining traction across various domains. They promise to bring significant improvements in people's lives, accelerating knowledge discovery and innovation. However, recently, there is an increasing concern regarding the societal and ethical implications of applying such techniques. Particularly, the lack of *diversity*, *fairness*, and *transparency* in statistical-based AI systems [10] raises a call for responsible by design systems [21, 26] that consider diversity and inclusion aspects [7], and brings about new regulations, like the EU's "Right to Explanation" [8].

Recommender Systems (RS) are a class of AI techniques that statistically analyze large traces of human behavior in order to extract preference patterns, which are then used to assist people in taking decisions. RS find applications in a broad range of domain 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. The prevalence of these systems in our lives naturally gives rise to concerns on their trustworthiness.

In this paper, we present our ongoing work towards the implementation of a framework called INFER that aims to improve the trustworthiness of recommender systems by combining ideas from three distinct research directions: *explainability*, *fairness* and *user interaction*. Explanations are an important mechanism to promote trust, as they provide human-understandable interpretations of the inner working of the system. However, they alone cannot assure that the system exhibits certain desirable ethical traits, such as being non-discriminatory, diversity-aware, bias-free, which we collectively refer to as aspects of *fairness*. Moreover, we believe it is vital that the human remains in the loop, able to *interact* with, and control the behavior of the system in an approachable manner. Although there exist recent works that handle individually or in combinations the aforementioned research directions, to the best of our knowledge, no approach takes the ambitious, holistic view of researching the intersection of these directions.

We set out the following *objectives* for the INFER framework: (1) to provide explanations capable of conveying, among others, fairness aspects; (2) to allow the auditing of the system for fairness and the discovery of latent source of bias via explanations; (3) to semi-automatically intervene to ensure system fairness; and (4) to interact with users to facilitate and guide the aforementioned functionality. We note that while the majority of existing works focus on certain use-cases, the envisioned framework targets the needs of different *stakeholders*, such as consumers of recommendations, providers/producers of objects to be recommended, system owners,

*All authors contributed equally to this research.

UMAP '21 Adjunct, June 21–25, 2021, Utrecht, Netherlands

ACM ISBN 978-1-4503-8367-7/21/06…$15.00
https://doi.org/10.1145/3450614.3462238

and regulators/auditors, who may have a varying level of familiarity and expertise with the system and the underlying technologies. The long-term vision is to enable these different stakeholders to query, understand, and fix sources of bias and discrimination in recommendations, in an accessible, understandable, and transparent manner.

This paper is structured as follows. In Section 2, we discuss the state of the art and the topics we find more challenging for advancing research in this area. In Section 3, we describe the major components of the INFER platform, including technical insights for each of them. Finally, in Section 4, we discuss conclusions and future steps.

2 BEYOND STATE-OF-THE-ART

Fairness. By fairness, we typically mean lack of discrimination. Typically, the definitions of fairness are based on the values of one or more sensitive attributes, such as gender or race. In RS, research discerns between fairness concerns appropriate to consumers and providers [6], with various definitions proposed [20]. Methods for achieving fairness in RS can be distinguished as pre- (modifying the input to the RS, e.g., by performing database repair [25]), in- (modifying algorithms to produce fair recommendations, e.g., by removing bias using matrix factorization [35] or variational autoencoders [5]) and post-processing (modifying the output of the algorithm, e.g., [13]). When fairness with respect to consumers and to item providers is important, an approach is to negotiate the trade-off between fairness and accuracy and improve the balance of user and item neighborhoods in collaborative filtering techniques [6].

As this research field is still in its early stages, it lacks clarity and consistency, with each work introducing a new definition of fairness. Our work attempts to categorize possible fairness concerns for different stakeholders. Moreover, current research typically operates on the premise that the sources of discrimination and the potential harms are known beforehand. However, there may exist latent sources of bias not previously recognized. An exhaustive examination of combinations of object attributes is neither sufficient nor feasible. Instead, we target at introducing explanations to interactively guide a human into discovering unrecognized, latent biases and harms, via semi-automatic processes.

Explainability. Explanations make AI systems more transparent and trustworthy. A line of research is on proxy models that approximate AI system decisions with a simpler interpretable model, e.g., [22]. Another direction is counterfactual explanations that present examples where the opposite decision would be observed [12, 33]. For RS specifically, 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) [29]. Modern RS 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 RS outputs. Regarding assessment of explainability for recommendations, [30] proposed a set of evaluation criteria, such as efficiency, effectiveness, persuasiveness, transparency and satisfaction.

Current techniques focus on a single stakeholder, namely the end-user of the recommender (i.e., the consumer) and seek to explain/interpret a single outcome of the system: e.g., why was I given this recommendation? In contrast, we investigate approaches that are suitable for different stakeholders, and require explaining a set of system outcomes: e.g., why did my product not appear in the recommendations to these people [28]? Moreover, existing explainability approaches provide general-purpose explanations, and are unsuitable for fairness-specific explanations that should incorporate notions of sensitive attributes, protected groups, treatment/impact, etc. In contrast, we target at techniques that assess explanations based on their importance in understanding the cause of unfairness.

Interactive Visual Explanations. Big data presents many challenges related to the rendering and presentation of large numbers of data points. Sampling techniques return approximate results [19], while aggregation techniques compute statistical summaries [11]. Recent work guides users with visualization recommendations [16], by considering user preferences and past behavior. Interactiveness in RS is typically served via visualization approaches and interactive explanations. Visualization approaches offer additional information to the user about the recommendations received, via histograms presenting rating distributions [9], tag clouds summarizing product content [32], and interactive graphs showing neighborhoods [14]. Our purpose of interactiveness and visualization is to improve or guide the explanations provided. [7] offers design guidelines to assess explainable user interfaces.

Although there exist several visualization methods and tools for a variety of datasets and algorithms, including AI systems, they focus on either presenting simplistic explanations to end-users or on trying to visualize complex models for experts. Further, the dimensions of fairness, automatic bias factors discovery and user-interactive fairness corrections and explanation are absent from such systems. INFER proposes new visual representations of aspects of fairness in RS and new visual operations for presenting rich explanations in an interactive manner.

3 INFER FRAMEWORK

This section presents the envisioned INFER framework. First, we present the objectives that prescribe the development of the proposed methods, then we described the proposed framework as a fully conceived and implemented platform supporting the above objectives. Finally, we discuss the expected impact of INFER.

3.1 Objectives

Objective 1: Fairness-aware explanations. The goal is to increase trust by providing explanations that convey fairness aspects. Examples of such explanations to consumers of recommendations and to product providers, respectively, are: "The system is gender-blind. A similar person like you but with opposite gender would get the same recommendations.", and "The system satisfies provider exposure fairness. Your products are recommended as frequently as those of other providers."

Objective 2: Explanation-based fairness audit. The goal is to increase the transparency of the system by examining whether fairness concepts are violated. Explanations can express a violation of fairness, e.g., when explanations as those in *Objective 1* cannot be provided. But more importantly, explanations can help identify underlying causes. For example, assume that the products of a tour operator are only being recommended to white tourists. An appropriate explanation engine might provide additional insight, e.g., these recommendations are better explained by values of the income attribute, indicating thus a bias in data (whites are over-represented in certain income groups). Going further, we envisage a system that efficiently and effectively navigates the space of possible explanations for a set of observations, and semi-automatically detects violations of fairness, that could not be manually defined or even recognized by experts.

Objective 3: Explanation-driven fairness by design. As per *Objective 2*, explanations may uncover causes of unfairness, and thus provide valuable feedback for the system owner in how to address them. For example, additional data or a calibration approach might be required to mitigate the unfair treatment of the tour operator. The challenge here is how to decide on the appropriate fairness intervention approach (pre-, in-, post-processing) based on fairness-aware explanations. Another complicating aspect is when multiple interventions from different stakeholders are required. The challenge is how to balance recommender effectiveness, while upholding multiple dimensions of fairness.

Objective 4: Interactive explanations. Explanations provide an intuitive, human-friendly way of understanding the inner workings of complex systems. Most approaches, however, tend to not close the loop, i.e., do not offer user control and feedback. Our objective is to design visual and interactive methods for users to control, refine, adjust the explanation received, via means of follow-up explanations, e.g., "But why was I not recommended this product?", and adjustments on fairness definitions and parameters, e.g., "Consider instead this group of people to be protected."

3.2 Research Directions

INFER proposes a holistic framework for supporting user-interactive, fairness-aware explanations on RS. The architecture, depicted in Figure 1, presents the core components that encapsulate the research to be performed, and which we detail next.

Explanation Engine. Our work starts off from previous work on the field [7], in order to identify suitable models to cover a wide range of stakeholders, with diverse roles, technical knowledge, and explanation needs from a recommender system. Specifically, we investigate how the explanation algorithms and interfaces in recommender systems can significantly achieve the explainability objectives such as interpretability and scrutability, which are not yet well explored in previous research. This will also lead to novel techniques for interpreting complex recommendation models.

An important research activity is to investigate how to adapt popular paradigms, such as proxy models [22], and counterfactual explanations [33], so that they can convey fairness concepts. In our work so far [12], we have developed a mechanism to generate counterfactual explanations for recommendations produced from arbitrary, opaque, and complex models. To explain a single recommender output, it may be sufficient to consider explanations that involve the sensitive attributes. However, to explain a set of outcomes, novel techniques beyond the state-of-the-art are necessary.

Fairness Audit Engine. This component serves two purposes, to explain why a specific fairness concept is violated, and to identify potential latent fairness violations. For the first, the goal is to determine the most important cause for a specific unfairness incident. Specifically, we will devise techniques that allow auditing for fairness based on explanations, an unexplored research direction. A large space of possible explanations will be considered, e.g., alternative local models, counterfactuals. Alternatives will be ranked based on the degree of unfairness they can explain (e.g., how likely it is to observe a specific counterfactual example that proves unfairness) and their suitability for different stakeholders (to be determined via user studies).

For the second purpose, we aim to extend our ideas in order to semi-automatically, and interactively detect unrecognized, latent fairness violations and their sources. To the best of our knowledge, no previous work has considered this direction. For this purpose, we assume that the end-user provides additional data features. Then the engine explores the space of explanations for a series of observations to identify potential instances of unfairness.

At the core of this engine, are algorithms that explore alternative explanations and enable stakeholders to audit for fairness even without having a fixed fairness concern. We will leverage our work in data exploration [31], and in auditing techniques [4], as well as recent work in the field [15].

Fairness Correction Engine. This component will use explanations to design and develop algorithms that semi-automatically correct fairness. In addition, we focus on manual fairness correction methods by providing to the end users a series of configurations that consider a number of fairness definitions, like statistical parity and equalized odds. We aim to extend both pre-processing and post-processing methods for ensuring fairness in order to allow fairness correction at the training data level or the ranked outputs level. Semi-automatic methods will be based on counterfactual fairness [15], while manual fairness correction methods will be able to exploit the whole range of the fairness audit engine, providing the end user a series of configurations w.r.t. candidate fairness dimensions, as well as state of the art fairness definitions.

Our focus will mainly be on extending post-processing fairness methods, so as to develop model-agnostic methods, ensuring their easy adoption. However, pre-processing methods will also be considered in order to allow fairness correction in the level of the training data, in cases this is required by certain stakeholders an scenarios. We will exploit our previous works on measuring popularity bias in collaborative filtering data [4], on how to ensure long term fairness in recommenders using variational autoencoders and other collaborative filtering models [3, 23, 27], and on how to balance effectiveness with fairness for consumers and providers [24].

Interactive Visualization Engine. This component will provide the end-user interface and the visual methods for users to visualize explanations and apply different explanation operations. These will cover the whole pipeline of: selecting and configuring explainability algorithms; exploring intermediate results and

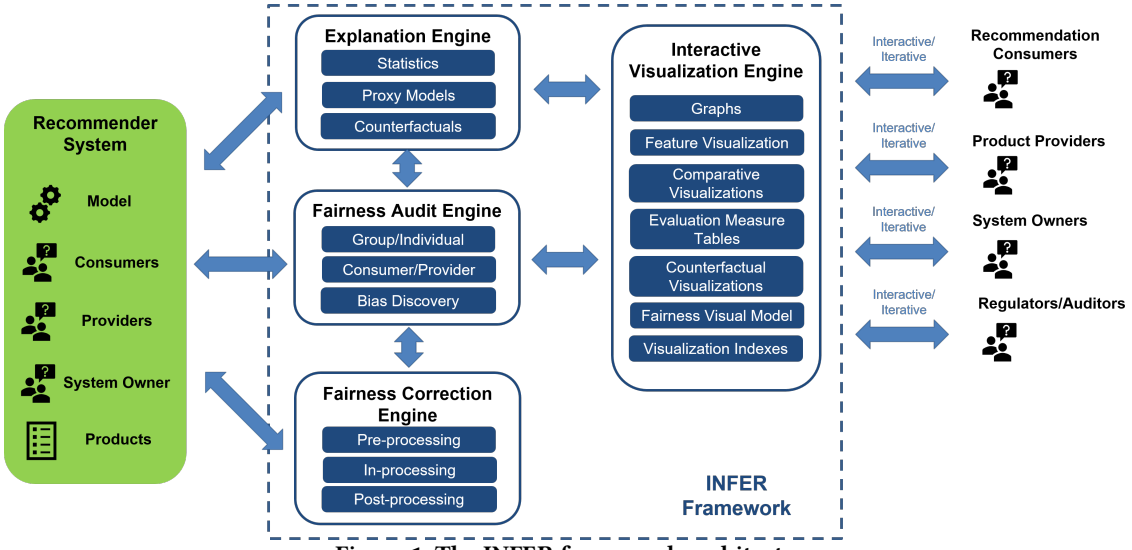

Figure 1: The INFER framework architecture.

re-configuring/tuning them; comparing the results of different explainability algorithms; measuring and ranking the fairness bias in fairness dimensions identified by the explainability models; selecting such dimensions for fairness correction; assessing fairness correction results.

We will study how explanation of fairness can be visually represented in the form of charts, such as traditional scatters or histograms, as well as counterfactual charts, feature visualizations, comparative charts, precision measure tables and dependency graphs. We will propose techniques for the iterative interaction of the user, i.e., how the user can interactively apply a series of visual operations that reveal and correct different aspects of fairness. We will base our work on our previous works of visualization tools ([1],[18], [2], [17]), considering also recent works on fairness-aware visualization, like the Silva tool [34]. The above visualization interfaces will provide a rich set of facilities, allowing expert users to delve into details of models and results, and use the obtained insights to properly configure fair recommendation models.

3.3 Expected Impact

INFER examines an interdisciplinary topic between societal studies and computer science. It develops novel algorithmic methods on explainability and visualization to leverage principles, such as transparency, trust, fairness and ethics into the functioning and the result of RS. Thus, INFER is expected to have impact in three directions: scientific, societal and commercial.

Scientific. INFER aims to interpret the inner representations and decisions of black-box models in RS and lay the foundations for such systems that can be audited and "tested for explainability" by design. The expected impact will affect the scientific community in the short and long term but will also open opportunities for societal and economic impact in the long term. Our goal is to draw the public's and researchers' attention to the potential societal risks that RS may bring and set the ground for a new generation of responsible recommenders by design.

Societal. User understanding, trust, transparency, usability and fairness are our core objectives. The identification of latent or complex sources of bias in existing RS and datasets will further abet towards discrimination elimination in societies. Further, INFER recognizes as stakeholders, with the right to the above principles, not only consumers, but also product providers, handling potential biases for both ends. It will also enable the auditing of RS systems by stakeholders that may only have a minimum level of understanding of the underlying technologies and concepts.

Commercial. The above societal opportunities can directly lead to commercial gains, since transparent, trustworthy and fair RS are expected to increase the engagement of both customers and vendors, and thus increase commercial exploitation/profits. INFER focuses on the user interactive explainability and fairness of RS, which comprise a category of AI systems widely used in a plethora of application areas. Further, at the core of the proposed research lies the effective integration of explainability and fairness methods and their exploitation via novel, interactive visualization tools.

4 CONCLUSION

In this paper, we have presented our vision on jointly researching the fields of explainability, fairness and user interactive visualization, towards enhancing the transparency, trust, fairness and user experience on recommender systems. Based on our previous work on the fields of statistically analyzing large data sets, identifying source of bias, and visualizing complex relationships, we identify existing gaps and challenges and we propose the INFER framework for addressing these challenges.

ACKNOWLEDGMENTS

This research has been co-financed by the Hellenic Foundation for Research and Innovation for the project *VisualFacts* (#1614) - 1st Call of Research Projects for the support of post-doctoral researchers, as well as by the EU project XMANAI (Explainable Manufacturing Artificial Intelligence, H2020-ICT-38-2020, GA ID 957362).

REFERENCES

[1] Nikos Bikakis, John Liagouris, Maria Krommyda, George Papastefanatos, and Timos K. Sellis. 2016. graphVizdb: A scalable platform for interactive large graph visualization. In *ICDE*.

[2] Nikos Bikakis, Stavros Maroulis, George Papastefanatos, and Panos Vassiliadis. 2021. In-situ visual exploration over big raw data. *Inf. Syst.* 95, 101616. https://doi.org/10.1016/j.is.2020.101616

[3] Rodrigo Borges and Kostas Stefanidis. 2019. Enhancing Long Term Fairness in Recommendations with Variational Autoencoders. In *MEDES*.

[4] Rodrigo Borges and Kostas Stefanidis. 2020. On Measuring Popularity Bias in Collaborative Filtering Data. In *BigVis*.

[5] Rodrigo Borges and Kostas Stefanidis. 2020. On Mitigating Popularity Bias in Recommendations via Variational Autoencoders. In *SAC*.

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

[7] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How should I explain? A comparison of different explanation types for recommender systems. *Int. J. Hum. Comput. Stud.* 72, 4 (2014).

[8] 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.

[9] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *CSCW*.

[10] Aylin Caliskan Islam, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora necessarily contain human biases. *CoRR* abs/1608.07187 (2016).

[11] Uwe Jugel, Zbigniew Jerzak, Gregor Hackenbroich, and Volker Markl. 2016. VDDA: automatic visualization-driven data aggregation in relational databases. *VLDB J.* 25, 1 (2016), 53–77.

[12] Vassilis Kaffes, Dimitris Sacharidis, and Giorgos Giannopoulos. 2021. Model-Agnostic Counterfactual Explanations of Recommendations. In *UMAP*.

[13] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2018. Recommendation Independence. In *FAT*.

[14] Bart P. Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. 2012. Inspectability and control in social recommenders. In *RecSys*.

[15] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *NIPS*.

[16] Yuyu Luo, Xuedi Qin, Nan Tang, and Guoliang Li. 2018. DeepEye: Towards Automatic Data Visualization. In *ICDE*.

[17] Stavros Maroulis, Nikos Bikakis, George Papastefanatos, and Panos Vassiliadis. 2021. RawVis: A System for Efficient In-situ Visual Analytics. In *SIGMOD (Demo)*.

[18] Stavros Maroulis, Nikos Bikakis, George Papastefanatos, Panos Vassiliadis, and Yannis Vassiliou. 2021. Adaptive Indexing for In-situ Visual Exploration and Analytics. In *DOLAP*.

[19] Dominik Moritz, Danyel Fisher, Bolin Ding, and Chi Wang. 2017. Trust, but Verify: Optimistic Visualizations of Approximate Queries for Exploring Big Data. In *CHI*.

[20] Evaggelia Pitoura, Georgia Koutrika, and Kostas Stefanidis. 2020. Fairness in Rankings and Recommenders. In *EDBT*.

[21] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. 2021. Fairness in Rankings and Recommendations: An Overview. *CoRR* abs/2104.05994 (2021). https://arxiv.org/abs/2104.05994

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

[23] Dimitris Sacharidis. 2019. Top-N group recommendations with fairness. In *SAC*.

[24] Dimitris Sacharidis, Kyriakos Mouratidis, and Dimitrios Kleftogiannis. 2019. A Common Approach for Consumer and Provider Fairness in Recommendations. In *RecSys Late-Breaking Results*.

[25] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. 2019. Interventional Fairness: Causal Database Repair for Algorithmic Fairness. In *SIGMOD*.

[26] Julia Stoyanovich, Serge Abiteboul, and Gerome Miklau. 2016. Data Responsibly: Fairness, Neutrality and Transparency in Data Analysis. In *EDBT*.

[27] Maria Stratigi, Jyrki Nummenmaa, Evaggelia Pitoura, and Kostas Stefanidis. 2020. Fair sequential group recommendations. In *SAC*.

[28] Maria Stratigi, Katerina Tzompanaki, and Kostas Stefanidis. 2020. Why-Not Questions & Explanations for Collaborative Filtering. In *WISE (Lecture Notes in Computer Science)*.

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

[30] Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems - Methodological issues and empirical studies on the impact of personalization. *User Model. User-Adapt. Interact.* 22, 4-5 (2012), 399–439.

[31] Georgia Troullinou, Haridimos Kondylakis, Kostas Stefanidis, and Dimitris Plexousakis. 2018. Exploring RDFS KBs Using Summaries. In *ISWC*.

[32] Jesse Vig, Shilad Sen, and John Riedl. 2009. Tagsplanations: explaining recommendations using tags. In *IUI*.

[33] 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).

[34] Jing Nathan Yan, Ziwei Gu, Hubert Lin, and Jeffrey M. Rzeszotarski. 2020. Silva: Interactively Assessing Machine Learning Fairness Using Causality. In *CHI*.

[35] Ziwei Zhu, Xia Hu, and James Caverlee. 2018. Fairness-Aware Tensor-Based Recommendation. In *CIKM*.