Fairness and Transparency in Recommendation: The Users' Perspective

Nasim Sonboli*
Jessie J. Smith*
nasim.sonboli@colorado.edu
jessie.smith-1@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

Florencia Cabral Berenfus
Florencia.CabralBerenfus@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

Robin Burke
Robin.Burke@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

Casey Fiesler
casey.fiesler@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

ABSTRACT

Though recommender systems are defined by personalization, recent work has shown the importance of additional, beyond-accuracy objectives, such as fairness. Because users often expect their recommendations to be purely personalized, these new algorithmic objectives must be communicated transparently in a fairness-aware recommender system. While explanation has a long history in recommender systems research, there has been little work that attempts to explain systems that use a fairness objective. Even though the previous work in other branches of AI has explored the use of explanations as a tool to increase fairness, this work has not been focused on recommendation. Here, we consider user perspectives of fairness-aware recommender systems and techniques for enhancing their transparency. We describe the results of an exploratory interview study that investigates user perceptions of fairness, recommender systems, and fairness-aware objectives. We propose three features – informed by the needs of our participants – that could improve user understanding of and trust in fairness-aware recommender systems.

CCS CONCEPTS

• **Human-centered computing → User studies**; **User centered design**; • **Information systems → Personalization**; **Recommender systems**.

KEYWORDS

fairness, transparency, qualitative study, recommender systems, explanation

ACM Reference Format:
Nasim Sonboli, Jessie J. Smith, Florencia Cabral Berenfus, Robin Burke, and Casey Fiesler. 2021. Fairness and Transparency in Recommendation: The Users' Perspective. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21), June 21–25, 2021, Utrecht, Netherlands.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3450613.3456835

1 INTRODUCTION & BACKGROUND

In recent years, the penetration of algorithmic systems into many realms of society has raised concerns that such systems may distribute benefits and harms unfairly across individuals and groups. This is particularly evident where high stakes decisions are made, ones that have significant impact on individuals' lives and livelihoods. Personalized recommender systems are a class of algorithms that learn from past user preferences in order to predict future interests of users and provide them with suggestions tailored to their tastes. In a personalized system, the objective of the recommendation algorithm is to accurately represent users' interests. However, in recent years, non-accuracy objectives such as fairness have become more common. This is because recommender systems are increasingly being employed in higher-stakes areas such as employment and financial services. Moreover, the user – that is, the consumer of the recommendation – is often not the only stakeholder in a recommender system. As a result, researchers have begun to explore how to ensure that a recommender system distributes its benefits fairly within and across different stakeholder groups, such as users and item providers [4, 6, 19, 28, 34, 46, 54]. In this paper, we describe findings from an exploratory interview study in which we elicited folk theories about recommender systems, as well as opinions about how fairness might be incorporated into such systems, and what forms of transparency around these algorithms are most effective for the users.

1.1 Folk Theories to Guide Explanation Design for Fair Recommendations

One method to collect and incorporate users' needs into a technical system is to examine their *folk theories* of that system [22, 31]. Folk theories in Human-Computer Interaction (HCI) are defined

[*] Both authors contributed equally to this research.

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

ACM ISBN 978-1-4503-8366-0/21/06…$15.00
https://doi.org/10.1145/3450613.3456835

as "intuitive, informal theories that individuals develop to explain the outcomes, effects, or consequences of technological systems" [14]. Folk theories can help us understand how users perceive the systems that they are interacting with [12, 20, 39], including uncovering inaccurate knowledge that users might have about a system in order to intervene on misguided or even harmful behavior on the platform [13, 20, 39]. These theories, accurate or not, shape the nature of user interaction and experience [21]. Therefore, it is important to broaden the education of users about the inner-workings of these systems as they lead to a better understanding of AI and web platforms [20, 35].

We argue that in order to design for transparency in a fairness-aware recommender system, we must first elicit feedback from the user. Folk theories are one means to recognize where users might have inaccurate knowledge about the system and its objectives. We propose to use explanations as a way to effectively educate users about these gaps in their knowledge, which might include fairness objectives in a recommendation system. Using explanations to educate in this way can help users create more accurate mental models of these systems, which can impact their actions on a platform, and help to inform future designs [26].

1.2 Fairness Objectives in Recommendation

Fairness in recommendation is generally assessed through two factors: representation and accuracy. Most of the proposed methods try to ensure parity in either of these factors among all user groups (e.g. women vs. men) or item groups (e.g. popular vs. unpopular). Parity is sought to avoid under-representation or over-representation of items/user subgroups, and to avoid under-estimation or over-estimation of accuracy for any item or user group [4, 6, 18, 19, 28, 45–47, 53, 54], as these issues could lead to discriminatory or unjust consequences for certain users.

Since recommender systems are multistakeholder systems, each stakeholder group may have different fairness needs. The key stakeholders in a recommendation system are the consumers (those who receive the recommendations, who we will also refer to as *users*), the providers (those who provide items that will be recommended), and the system/platform (that hosts the recommendations) [7]. Fairness can be defined and sought for any of the stakeholders involved. In this paper we focus on one aspect of fairness-aware recommendation: *provider fairness.* In our interviews, we frame the objective of provider fairness as a method that increases the representation of items that are consistently not being recommended.

1.3 Transparency & Explanations

One of our key areas of interest in this study is the question of how to design for transparency in the context of fairness-aware algorithms. Transparency has become an essential feature of algorithmic design, especially in application areas that are socially sensitive due to (1) legal requirements for transparency such as the GDPR [24, 41]; (2) the need to build users' trust in a system [11, 29, 38]; (3) the need for error detection to help mitigate bias and discrimination in a system [23, 37, 48, 49]; (4) helping further public adoption of new technologies [1, 3, 17, 25, 30]; and/or (5) creating an environment of accountability for the platforms that host the algorithms [1, 15].

We believe that transparency only increases in importance for fairness-aware systems [32]. Because of the complex and contested nature of fairness as a concept [36], it is not enough to claim that a system is fair: users need to understand the specifics of the fairness objectives that have been encoded into it. It is well recognized that users of recommender systems benefit from explanations of the recommendations that they receive [10, 27, 49]. Because fairness-aware recommendation algorithms influence the recommendations users receive in a way that responds to fairness goals, we argue that it is particularly important that recommender systems be transparent about their fairness objectives.

Transparency for fairness-aware recommendations is different from general transparency in recommendation algorithms (e.g., as discussed in [43, 49]) in that the goal is not only to explain how the system works, but also to provide clarity for what its fairness goals are, the motivations for these goals, and how they might impact different stakeholders. However, explanations of fairness-aware algorithms must be done with care; the *ways* in which fair algorithms are explained to users have been proven to heavily influence their trust, adoption, and understanding of these systems [9, 52]. Since explanations in a fairness-aware system impact users' trust of and actions within the system, it is especially important to ensure that these explanations are created effectively and carefully, with the users' needs in mind.

While explanation has a long history in recommender systems research [10, 27, 49], there has been little work that addresses explaining systems that involve a fairness objective. Similarly, while previous work in other branches of AI has explored the use of explanations as a tool to increase fairness (for example, by reducing unwanted bias or discrimination) [2, 16, 51], this work has not focused on recommendation specifically.

This paper contributes findings about participants' folk theories of recommendations and fairness, as well as their opinions about transparency and explanation. Based on these findings, we take the position that explanations ought to be used as a means to educate users about algorithmic objectives, especially in fairness-aware systems, and suggest three features that should be included in effective explanations of fairness-aware recommender systems.

2 METHODS

In this work, we use interviews as a way to explore the stories and experiences of a sample of users of recommender systems. Though some previous work has stated that user input for algorithmic design may not be helpful for engineers, due to users' potential lack of knowledge of the system [42], we argue that it is precisely this lack of knowledge that can be beneficial for the design and deployment of *explanations* in systems. Understanding what the users do *not* know about a system can help the engineer pinpoint where users might need more education about the platforms they are interacting with, and how an explanation can meet those needs.

As an initial exploratory study, we recruited a convenience sample of 30 undergraduate and graduate students from a large state university in the United States, all of whom had prior exposure to recommender systems (e.g., on e-commerce or streaming platforms such as Amazon or Netflix). Our participants included 18 women and 12 men, with an age range of 18 to 32. 15 participants were studying Computer Science, Information Science, or related fields, and 15 were pursuing other majors. Participants were recruited via social media postings, a university announcement board, and advertisements in classes, and were compensated $10 for their time; this study was approved by our institution's IRB. Though our sample of young internet users provides a good foundation for understanding opinions and concerns that ordinary recommendation consumers might have about fairness, we recognize that it is limited in scope and we do not suggest that our findings are generalizable to a broader population.

We conducted face-to-face, semi-structured interviews with participants that began with a discussion of their past experiences with recommender systems and their folk theories of recommendation algorithms and fairness objectives. We asked them to explain their current knowledge about how these systems work, and what fairness meant to them in this context. For the latter half of the interviews, we used the Kiva crowd-sourced microlending platform (explained in more detail below) as a case study to explore the ways that explanations can help or harm the participant's understanding of a fairness-aware system.

Following interview transcription, researchers conducted thematic analysis [5]. Three independent analysts conducted open-coding on a subset of the interviews using MAXQDA software, and then met to confer on, synthesize, and finalize a set of themes that best captured the insights gathered from our participants. We conducted a detailed analysis of each theme, resulting in the overall story that our data tells, as described in detail in the Results and Discussion sections.

2.1 Kiva as a Case Study

Kiva is an organization seeking to enhance financial inclusion in the world through microlending. Their platform provides a space for those who are seeking loans (borrowers) to get funded by those who are seeking to supply capital (lenders). Kiva's mission emphasizes equitable access to capital for all borrowers, who are individuals without access to traditional forms of capital, to improve their living situations. If Kiva implemented a recommender system, the providers of the recommended content would be the borrowers, and the consumers of the recommendations would be the lenders [8]. Naturally, a recommender system in this context raises an important fairness concern. If borrowers are recommended inequitably, the system could be contributing to an inequitable pattern of resource distribution on its platform. We used Kiva as a case study both to have a concrete example to ensure consistency across participants, and because Kiva is a real-world example that contains fairness concerns. In our interviews, we asked the participants to consider what helpful explanations might look like in Kiva's platform if fairness-aware recommendation were implemented.

3 RESULTS

In this section, we describe the main findings from our exploratory interviews with users of recommender systems. We begin by explaining some of the gaps in participants' knowledge about recommender systems and fairness objectives, as gathered from participant folk theories. Then, we explore the ways in which participants indicated that explanation design of a fairness-aware system could help or harm their understanding and trust of the system. Participant quotes are indicated by anonymized participant numbers; participants 1-15 were studying CS or adjacent fields, and 16-30 were in non-technical fields.

3.1 Folk Theories of Recommender Systems

Some participants expressed that the "black box" nature of recommendations made it difficult for them to learn how a recommendation system might create personalized recommendations for them. For example, as P5 said, "recommender systems in most cases are pretty unknown to the user."

However, in general, participants had somewhat of an understanding for how recommender systems worked – particularly in terms of how much data was being collected on them. Some participants' folk theories aligned well with accuracy-based recommendation algorithms, such as collaborative filtering algorithms like User-KNN or Item-KNN.

> "So if like a lot of people buy, you know, a hammock, then if a large subset of those people who bought a hammock, also bought a sleeping bag, it might recommend the sleeping bag to you." – P21

3.2 Folk Theories of Fairness Objectives

When asked about how recommendations could be unfair to users, very few participants indicated that they had ever thought of provider fairness for recommendations. This raised a concern that there might be a lack of communication between the system and the user when it comes to issues of provider fairness in recommendation.

After a short discussion about fairness objectives and the impact that a recommender system might have on the providers, many participants indicated that they thought provider fairness was important to them in a recommender system, as expressed by P18.

> "I don't think I've thought about it as much from a seller's perspective, but I can see like, you know, these platforms are made for more than just [users]. They're made for [providers] as well. So like people who sell things on Amazon or make music and put it on Spotify, it's like, yeah, you're probably inherently at a disadvantage to people who are already big or are already in favor of the algorithm." – P18

However, the majority of participants were still not entirely sure how fairness goals might impact their recommendations. Thus, we began to explore different ways that a recommendation platform could explain this impact to users in an educational and empowering way.

3.3 Explaining Fairness Goals to Users

Throughout the interviews, most participants indicated that they would want to see the fairness goals of an organization described to the users in some way, either through a short explanation or an entire page, as described by P8.

> "[Organizations] should have [fairness goals] somewhere I could find it like the little 'about us', like 'learn more about our corporation' tab where they would

> explain 'these are our moral values, these are what we prioritize'." – P8

One participant thought that fairness goals were best incorporated into the UX of the platform, as long as the language did not unintentionally manipulate users.

> "Or like work [fairness goals] into their platform somehow. Like how Spotify does the 'New Music Fridays'... I think that there are opportunities to weave that into places on a platform but not necessarily blanket it across so the user doesn't feel like they have choice." – P29

This raises an important concern that is not new in the field of explanation: the concern that explanations, if not designed carefully, could be used as a tool to manipulate users.

3.4 Fairness as a User Choice

One outcome of fairness-aware recommendation is to "nudge" users into choosing items that they might otherwise not. The intention of this kind of system is to encourage the user to choose the items that meet the fairness objective of the system itself, and it is precisely this intention that might underlie some participants' concerns with being "manipulated". In fairness-aware recommender systems, *coercing* users into choosing a "fair" recommendation – which might not be 'fair' for everyone – could be misleading. This concern was expressed amongst our participants, particularly because fairness definitions are often disputed, and what might be considered *fair* for some, could seem *unfair* for others [44]. In order to remedy these concerns, several of our participants indicated that they would prefer to be informed about the existence of fairness-aware recommendations so that the user could make the choice of fairness or personalization for themselves, as expressed by P22.

> "[Fairness-aware recommendation] is manipulative in some sort of way. I think the best thing that they can do would be to give a short explanation that they changed [the personalized recommendation algorithm] and then kind of show the whole list and not point out any others in the list... allowing an unbiased choice from the viewer." – P22

If manipulation was a concern amongst participants, then how should explanations be designed to mitigate that concern? Further, how might explanations instead be designed to foster trust and communication between a system and its users? In the next section, we propose that these concerns can be alleviated through explanations if they are used effectively as a tool for education.

3.5 Explanations as Education

One prominent theme that emerged throughout the interviews was that *transparency as a means for education was essential.* Whether an algorithm uses fairness objectives or not, participants expressed that they needed to be educated about why they were being recommended the items that they were. This point was succinctly summed up by P21 who stated: "the more transparency, the better."

Many participants indicated that better design practices could help promote fair treatment for providers and relay this fair treatment back to the recommendation consumers. In order to ground participants in a specific platform and discuss concrete explanation designs, we used Kiva as a case study. Specifically, we asked the participants how they thought transparency could be designed when fairness goals were incorporated in a system like Kiva. We gathered feedback about the benefits and flaws of explanation designs that were minimal and global, versus detailed and specific.

Several participants were in favor of including one global explanation for all fairness-aware recommendations. Some indicated that a single explanation could still leave room for the system to include more in-depth explanations about a specific group of providers that the system is optimizing for fairness. For example, many participants expressed that they would like to know more about why Kiva borrowers from a specific geographic area are underfunded on the platform, and how choosing to lend to people from that area could have a positive social impact.

> "I like [longer explanations], where you can scroll underneath and get like a broader perspective of what's going on in the whole region, I would like that." – P5

Another alternative was to provide a separate explanation for every individual recommendation, on the recommended item itself. Many participants expressed that if explanations were specific to each recommendation, the manner in which they were formatted played a big role in their utility to the consumer. P30 described that for explanations to be effective and informative, the *"algorithms should be explained to you in a simple sentence structure... short, concise and in digestible pieces".* They added that explanations should be accessible and inclusive for all audiences, regardless of their level of understanding of algorithms.

In summary, we found that when it comes to transparency for fairness-aware recommendations, design decisions matter. Transparency should give users greater agency to understand how provider fairness might make their recommendations different than personalization. Further, when recommendation fairness is explained to the consumer, it should be explained in a way that is accessible, understandable, specific, and concise.

4 DISCUSSION & CONCLUSIONS

In this study, we explored what users of recommender systems understand about personalized and fairness-aware recommendations. Through semi-structured interviews, we used folk theories as an effective method for pinpointing gaps and inaccuracies in user knowledge of a system. This formed our basis for understanding that transparency in recommender systems ought to serve as a means of education for users, especially in a fairness-aware system where the objectives are often unknown or misunderstood by the user. Taking into account the feedback provided by the interview participants, we provide the following suggestions for features that should be included in effective explanation design for fairness-aware systems.

1. *Explanations should define the system's fairness objective for users.* For example, an explanation could educate the user about the impact that a fairness-aware recommendation might have on the item providers.
2. *Explanations should not nudge/manipulate users into making a decision, even if the goal is fairness.* For example, an explanation could educate users about the existence of a

fairness-aware recommendation while still allowing the user to decide if they prefer personalization instead.

(3) *Explanations should disclose the motivation for using fairness as a system objective*. For example, an explanation could educate users about the fairness concerns of the system, or the values of the organization and why they chose fairness as an objective.

These three features were compiled based on the specific gaps that our participants had in their knowledge about fairness-aware recommender systems. Specifically, the folk theories that participants shared on these systems led us to believe that there might exist a major gap in communication towards users about recommender systems' objectives. These features also summarize participants' most frequent desires for what they wished explanations could teach them about the system in order for them to understand and trust it more. It is also interesting to note that whenever participants were educated about provider fairness concerns in recommendations, they generally showed interest in the option of using a system that had fairness as an objective.

However, we note some limitations in this study's exploration of participants' opinions around fairness, transparency and explanation. Firstly, presenting examples of provider fairness in a face-to-face interview setting could have resulted in acquiescence response bias, or a tendency for participants to frame their views in ways that they believed was expected from or valued by the interviewer [33]. Moreover, it is possible that social desirability response bias could have influenced participants' expressed interest in provider fairness, when discussing Kiva as a nonprofit organization seeking to increase financial access globally [40, 50]. It may also be the case that recommender systems that do not carry such high stakes for providers (e.g., a food delivery recommender system) might attract different levels of interest from users when it comes to provider fairness. Thus, we suggest that future work could further investigate similar questions by way of anonymous self-reports, or by assessing participant behavior in an experimental design.

With this work, we encourage organizations that are trying to incorporate fairness into the design of their recommender systems to seek input from their users when designing for transparency. If transparency design fails to consider the knowledge (or lack thereof) of the users, the platform runs the risk of losing user trust. Moreover, if fairness goals are explained to users in vague or ineffective ways, this can cause confusion or lead to misunderstanding of the goals of the system. In turn, when fairness goals are explained to the users effectively, they can gain agency, and be left feeling more empowered and knowledgeable about the platform and their impact while using it. Ultimately, we argue that in order to gain users' trust in and adoption of a fairness-aware recommender system, collaboration is *necessary*. Through effective communication, feedback, and transparency, recommender systems can contribute to a more empowering, fair, and trustworthy digital future.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under Grant No. 1911025.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–18.

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[3] Victoria Bellotti and Keith Edwards. 2001. Intelligibility and accountability: human considerations in context-aware systems. *Human–Computer Interaction* 16, 2-4 (2001), 193–212.

[4] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in Recommendation Ranking through Pairwise Comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2212–2220. https://doi.org/10.1145/3292500.3330745

[5] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[6] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. 2018. Balanced Neighborhoods for Multi-sided Fairness in Recommendation. In *Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 202–214. http://proceedings.mlr.press/v81/burke18a.html

[7] Burke, Robin. 2017. Multisided Fairness for Recommendation. In *Workshop on Fairness, Accountability and Transparency in Machine Learning (FATML)*. arXiv, Halifax, Nova Scotia, 5 pages. arXiv:1707.00093[cs.CY]

[8] Jaegul Choo, Changhyun Lee, Daniel Lee, Hongyuan Zha, and Haesun Park. 2014. Understanding and Promoting Micro-Finance Activities in Kiva.Org. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining* (New York, New York, USA) *(WSDM '14)*. Association for Computing Machinery, New York, NY, USA, 583–592. https://doi.org/10.1145/2556195.2556253

[9] Jason A Colquitt and Jerome M Chertkoff. 2002. Explaining injustice: The interactive effect of explanation and outcome on fairness perceptions and task motivation. *Journal of Management* 28, 5 (2002), 591–610.

[10] Dan Cosley, Shyong K. Lam, Istvan Albert, Joseph A. Konstan, and John Riedl. 2003. Is Seeing Believing? How Recommender System Interfaces Affect Users' Opinions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 585–592. https://doi.org/10.1145/642611.642713

[11] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-adapted interaction* 18, 5 (2008), 455.

[12] Michael A. DeVito, Jeremy Birnholtz, Jeffery T. Hancock, Megan French, and Sunny Liu. 2018. How People Form Folk Theories of Social Media Feeds and What It Means for How We Study Self-Presentation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173694

[13] Michael A DeVito, Jeremy Birnholtz, Jeffery T Hancock, Megan French, and Sunny Liu. 2018. How people form folk theories of social media feeds and what it means for how we study self-presentation. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[14] Michael A. DeVito, Darren Gergle, and Jeremy Birnholtz. 2017. "Algorithms Ruin Everything": #RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3163–3174. https://doi.org/10.1145/3025453.3025659

[15] Nicholas Diakopoulos. 2015. Algorithmic accountability: Journalistic investigation of computational power structures. *Digital journalism* 3, 3 (2015), 398–415.

[16] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608 [stat.ML]

[17] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. *International journal of human-computer studies* 58, 6 (2003), 697–718.

[18] Michael D Ekstrand, Robin Burke, and Fernando Diaz. 2019. Fairness and Discrimination in Recommendation and Retrieval. In *Proceedings of the 13th ACM Conference on Recommender Systems* (Copenhagen, Denmark) *(RecSys '19)*. Association for Computing Machinery, New York, NY, USA, 576–577. https://doi.org/10.1145/3298689.3346964

[19] Michael D. Ekstrand, Mucun Tian, Mohammed R. Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. 2018. Exploring Author Gender in Book Rating and Recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 242–250. https://doi.org/10.1145/3240323.3240373

[20] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I "like" It, Then I Hide It: Folk Theories of Social Feeds. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2371–2382. https://doi.org/10.1145/2858036.2858494

[21] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300724

[22] Muheeb Faizan Ghori, Arman Dehpanah, Jonathan Gemmell, Hamed Qahri Saremi, and Bamshad Mobasher. 2019. Does the User Have A Theory of the Recommender? A Pilot Study.. In *Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, RecSys '20*. Copenhagen, Denmark, 77–85.

[23] Jennifer Goetz, Sara Kiesler, and Aaron Powers. 2003. Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.* Ieee, 55–60.

[24] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine* 38, 3 (2017), 50–57.

[25] Shirley Gregor and Izak Benbasat. 1999. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly* (1999), 497–530.

[26] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining Collaborative Filtering Recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work* (Philadelphia, Pennsylvania, USA) *(CSCW '00)*. Association for Computing Machinery, New York, NY, USA, 241–250. https://doi.org/10.1145/358916.358995

[27] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining Collaborative Filtering Recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work* (Philadelphia, Pennsylvania, USA) *(CSCW '00)*. Association for Computing Machinery, New York, NY, USA, 241–250. https://doi.org/10.1145/358916.358995

[28] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Issei Sato. 2016. Model-based approaches for independence-enhanced recommendation. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. IEEE, IEEE, New York, 860–867.

[29] René F. Kizilcec. 2016. How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2390–2395. https://doi.org/10.1145/2858036.2858402

[30] René F Kizilcec. 2016. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2390–2395.

[31] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1–10.

[32] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology* 31, 4 (2018), 611–627.

[33] Mingnan Liu, Mingnan Liu, Frederick G. Conrad, Frederick G. Conrad, Sunghee Lee, and Sunghee Lee. 2017. Comparing acquiescent and extreme response styles in face-to-face and web surveys. *Quality & quantity* 51, 2 (2017), 941–958.

[34] Weiwen Liu, Jun Guo, Nasim Sonboli, Robin Burke, and Shengyu Zhang. 2019. Personalized Fairness-Aware Re-Ranking for Microlending. In *Proceedings of the 13th ACM Conference on Recommender Systems* (Copenhagen, Denmark) *(RecSys '19)*. Association for Computing Machinery, New York, NY, USA, 467–471. https://doi.org/10.1145/3298689.3347016

[35] Duri Long and Brian Magerko. 2020. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–16.

[36] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, Vol. 1170. New York, NY, USA, 1 pages.

[37] P J Phillips, Amanda C Hahn, Peter C Fontana, David A Broniatowski, and Mark A Przybocki. 2020. Four Principles of Explainable Artificial Intelligence (Draft). (2020).

[38] Pearl Pu and Li Chen. 2007. Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems* 20, 6 (2007), 542–556.

[39] Emilee Rader and Rebecca Gray. 2015. Understanding User Beliefs About Algorithmic Curation in the Facebook News Feed. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 173–182. https://doi.org/10.1145/2702123.2702174

[40] Donna M. Randall and Maria F. Fernandes. 1991. The Social Desirability Response Bias in Ethics Research. *Journal of business ethics* 10, 11 (1991), 805–817.

[41] Protection Regulation. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council. *REGULATION (EU)* 679 (2016), 2016.

[42] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. 2019. How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 99–106. https://doi.org/10.1145/3306618.3314248

[43] Rashmi Sinha and Kirsten Swearingen. 2002. The Role of Transparency in Recommender Systems. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems* (Minneapolis, Minnesota, USA) *(CHI EA '02)*. Association for Computing Machinery, New York, NY, USA, 830–831. https://doi.org/10.1145/506443.506619

[44] Jessie Smith, Nasim Sonboli, Casey Fiesler, and Robin Burke. 2020. Exploring User Opinions of Fairness in Recommender Systems. *Human-Centered Approach to Fair & Responsible AI Workshop at CHI '20* (2020), arXiv–2003.

[45] Nasim Sonboli and Robin Burke. 2019. Localized Fairness in Recommender Systems. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) *(UMAP'19 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 295–300. https://doi.org/10.1145/3314183.3323845

[46] Nasim Sonboli, Farzad Eskandanian, Robin Burke, Weiwen Liu, and Bamshad Mobasher. 2020. Opportunistic Multi-Aspect Fairness through Personalized Re-Ranking. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20)*. Association for Computing Machinery, New York, NY, USA, 239–247. https://doi.org/10.1145/3340631.3394846

[47] Harald Steck. 2018. Calibrated Recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 154–162. https://doi.org/10.1145/3240323.3240372

[48] Andreas Theodorou, Robert H Wortham, and Joanna J Bryson. 2017. Designing and implementing transparency for real time inspection of autonomous robots. *Connection Science* 29, 3 (2017), 230–241.

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

[50] Thea F van de Mortel. 2008. Faking It: Social Desirability Response Bias in Self-report Research. *Australian journal of advanced nursing* 25, 4 (2008), 40–48.

[51] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[52] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. 2020. *Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376813

[53] Lin Xiao, Zhang Min, Zhang Yongfeng, Gu Zhaoquan, Liu Yiqun, and Ma Shaoping. 2017. Fairness-Aware Group Recommendation with Pareto-Efficiency. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* (Como, Italy) *(RecSys '17)*. Association for Computing Machinery, New York, NY, USA, 107–115. https://doi.org/10.1145/3109859.3109887

[54] Sirui Yao and Bert Huang. 2017. Beyond Parity: Fairness Objectives for Collaborative Filtering. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., Red Hook, NY, USA, 2921–2930.