Toward Natural Language Mitigation Strategies for Cognitive Biases in Recommender Systems

Alisa Rieger
TU Delft
Van Mourik Broekmanweg 6
2628 CD Delft
a.rieger@tudelft.nl

Mariët Theune
University of Twente
Drienerlolaan 5
7522 NB Enschede
m.theune@utwente.nl

Nava Tintarev
TU Delft
Van Mourik Broekmanweg 6
2628 CD Delft
n.tintarev@tudelft.nl

Abstract

Cognitive biases in the context of consuming online information filtered by recommender systems may lead to sub-optimal choices. One approach to mitigate such biases is through interface and interaction design. This survey reviews studies focused on cognitive bias mitigation of recommender system users during two processes: 1) item selection and 2) preference elicitation. It highlights a number of promising directions for Natural Language Generation research for mitigating cognitive bias including: the need for personalization, as well as for transparency and control.

1 Introduction

Decision-making at an individual, business, and societal levels is influenced by online news and social media. Filtering and ranking algorithms such as recommender systems are used to support these decisions. Further, individual cognitive selection strategies and homogeneous networks can amplify bias in customized recommendations, and influence which information we are exposed to (Bakshy et al., 2015; Baeza-Yates, 2018).

Biased exposure to online information is known to accelerate extremism and the spread of misinformation (Hills, 2019). Ultimately, these undesirable negative consequences of information filtering diminish the quality of public discourse and thus can pose a threat to democracy (Bozdag and van den Hoven, 2015).

One strategy for bias mitigation would be to raise users' awareness of filtering mechanisms and potential cognitive biases. Approaches going one step further than creating awareness, actively nudge users in a direction of less biased information selection and diversification. Explanations and nudges for mostly non-expert users of recommender systems in the domains of news and social media have to be designed in a way that they are understood intuitively, e.g., using natural language (Liao et al., 2020).

To our knowledge, no previous work has summarized cognitive bias mitigation in the context of recommender systems. In this paper, we aim to identify research gaps and opportunities to improve natural language explanation interfaces that mitigate cognitive biases. We do this by providing an overview of approaches to mitigate cognitive bias of recommender system users in the domains of news and social media. We review the literature in the field and summarize ways of measuring bias and mitigation approaches for different biases in different contexts. We also consider how these occur at different stages of the recommendation process. In sum, we address the following research questions (RQs):

1. For which types of cognitive biases occurring among users of recommender systems exist validated mitigation approaches?
2. What are effective approaches to *measure* different types of bias?
3. What are effective approaches to *mitigate* different types of bias?
4. How are the mitigation approaches *evaluated*?

In the next section, we introduce the method used in our literature review. Then, in Section 3, we analyze the resulting papers and identify commonalities. We see that human bias mitigation using natural language generation in recommender systems is still under-explored despite explanations being successfully applied in the fields of persuasive technology and argumentation (Dragoni et al., 2020; Guerini et al., 2011). So, in Section 4 we take a constructive approach and discuss promising directions for natural language generation (NLG) research, before concluding in Section 5.

2 Methodology

To find relevant literature for this survey, we defined inclusion criteria as a search string which we ran through the databases Springerlink (http://link.springer.com) and ACM digital library (https://dl.acm.org) in July 2020. These two databases are established and comprehensive databases in the field of computer science, and support complex search strings. The search results were filtered by scanning Title, Abstract, and Discussion.

Inclusion criteria: Our search string covers four main concepts: **(1)** bias-related; **(2)** target-system-related; **(3)**; domain-related; **(4)** mitigation-related. The terms used for each concept are: **(1)** *(”cognitive bias” OR ”human bias” OR ”confirmation bias” OR ”availability bias” OR ”backfire effect” OR ”homophily” OR ”affinity bias” OR ”decoy effect” OR ”selective exposure” OR ”false consensus effect” OR ”saliency bias”) AND* **(2)** *(”recommender” OR ”recommendation”) AND* **(3)** *(”news” OR ”social media” OR ”search” OR ”information seeking”) AND* **(4)** *(”mitigat*” OR ”debiasing” OR ”reduce” OR ”explainable artificial intelligence” OR ”XAI” OR ”intelligent user interface” OR ”IUI” OR ”natural language”).* This search resulted in 257 hits.

Exclusion criteria: Papers are excluded if they do not: **a)** focus on recommender systems in the domains of news, social media, or search (40 excluded); **b)** do not propose a mitigation approach for human bias (137); **c)** do not present a user study (66); **d)** do not include measures of bias (5); **e)** we have no access to the full paper (5). These criteria lead to the exclusion of 253 papers, resulting in the four papers discussed in the remained of this paper (see Table 1). We observe that these papers do not cover linguistic solutions, but will later see that they still highlight promising areas for research in NLG.

3 Analysis

In this section we analyze and compare the four resulting papers based on five aspects which were chosen to answer the research questions: **(RQ1)** *Objective*: context and objective of the paper and *Bias*: type of cognitive bias investigated; **(RQ2)** *Measure*: approach for measuring bias; **(RQ3)** *Mitigation*: approach of bias mitigation; and **(RQ4)** *Evaluation*: evaluation of the mitigation approach and moderating factors.

(RQ1) *Objective and Bias*: To encourage diverse information and common ground seeking, Liao and Fu (2014) investigated the mitigation of selective exposure or the confirmation bias, which is the tendency to search for and select information which confirms previous beliefs and values, in online discussion forums. Graells-Garrido et al. (2016) researched the mitigation of confirmation bias and homophily, the tendency to have and build ties to similar individuals to oneself, with the intention to connect users with different opinions in social networks. Tsai and Brusilovsky (2017) studied the mitigation of homophily and position bias, occurring if the position influences the perceived value or utility of an item, in the context of a tool for conference attendees to connect to diverse scientists. Pommeranz et al. (2012) intended to design user interfaces for unbiased preference elicitation, which are needed for accurate recommendations. Preference elicitation describes the process of collecting user data to build an accurate user-model, based on which items are recommended. Thus, Pommeranz et al. (2012) investigate bias mitigation at an earlier stage in the recommendation process, than the other three reviewed studies. The authors list a number of possible biases that can occur during the stage of preference elicitation (but do not measure them): *framing* – presentation with positive or negative connotations influence the perceived value or utility of an item, *anchoring* – value of an initially encountered item influences the perceived value of a subsequently encountered item, and *loss aversion* – tendency to prefer avoiding losses to obtaining gains with the same value.

(RQ2) *Measure*: To measure bias, all of the studies compared the effect of an intervention with a baseline system on a set of metrics. For the three studies researching confirmation bias and homophily during item selection, the diversity of item selection or the degree of exploration of items was compared to the baseline (without bias mitigation) (see Liao and Fu, 2014; Graells-Garrido et al., 2016; Tsai and Brusilovsky, 2017). Diversity and degree of exploration were calculated on basis of the users’ clicking behavior and attributed values for each item, reflecting the aspects of interest in the study (e.g., position - pro/con, similarity of profile - high/low,..). For framing, anchoring, and loss aversion during preference elicitation, a quality score was calculated for each tested preference elicitation method. A high level of agreement between the system’s outcome preference model and the user-generated list of preferences resulted in a

	Bias	Objective	Mitigation
Liao and Fu, 2014	confirmation bias	viewpoint diversification of users in forum for political discussions	*Visual barplot*: indication of source position valence and magnitude to reduce the demand of cognitive resources
Graells-Garrido et al., 2016	confirmation bias and homophily	connecting users with diverse opinions in social networks	*Visual data portraits and clustering*: indication of own interests and opinions as data portrait to explain recommendations, and display of users with shared latent topics in interactive clusters to facilitate exploration
Tsai and Brusilovsky, 2017	homophily and position bias	help conference attendees to connect to diverse scientists via a social network	*Multidimensional visual scatterplot*: display of scientists' accademic and social similarity and highlights potential matches through color-coding
Pommeranz et al., 2012	framing, anchoring, loss aversion	designing user-centered interfaces for unbiased preference elicitation	*Multiple visual interface proposals*: virtual agent with thought bubble, outcome view (explore link between interests, preferences and outcomes), interest profiling, affective feedback,..

Table 1: Examined Bias, Objective, and Mitigation approach per paper

high quality score (see Pommeranz et al., 2012).

(RQ3) *Mitigation*: Liao and Fu (2014) displayed posts in the online forum in combination with a visual barplot which indicated position valence (pro/con) and magnitude (moderate/extreme) of the posts' authors to mitigate confirmation bias. The authors argue that freeing up cognitive resources can increase users capacity to assess viewpoint challenging information. They aimed to reduce the demand on cognitive resources by pre-evaluating and marking the author's position, with the intention that this would increase users' capacity to process information relating to the post's *content*.

Further, the explicit indication of author position information aimed at encouraging attention to diverse viewpoints and motivating users to select attitude-challenging information. Graells-Garrido et al. (2016) recommended diverse profiles with shared latent topics and displayed visualizations of the user's own data portrait in the form of word-clouds with interests and opinions to explain the given profile recommendations and mitigate confirmation bias and homophily. Profile recommendations were presented in the form of visual clusters of accounts with shared latent intermediary topics, from which the user could select accounts for exploration. This approach aimed to overcome cognitive dissonance produced by direct approaches of exposure to challenging information. The aim was to provide context to a given recommendations, both in form of the user's own data profile and the basis of a shared intermediary topic, to give the new connection a chance. Another approach to mitigate homophily in addition to position biases was chosen by Tsai and Brusilovsky (2017), who presented scientists as points in a two-dimensional scatterplot. The position of a point was calculated by social (co-authorship) and academic (publication content) feature similarity (0 - 100 %) between user and scholar. Meaningful feature combinations, defined by higher degrees of feature similarities, were highlighted through color-coding. This approach aimed to enable the presentation of more than one recommendation aspect, to guide conference attendee's attention to areas of scientists with meaningful feature combinations, and overall, to promote diversity of profile exploration. Pommeranz et al. (2012) propose input methods and interfaces for preference elicitation which result in equal mental preference model and system preference representation to achieve a mitigation of framing, anchoring and loss aversion biases. They investigated different methods of preference elicitation, such as rating with a nine point likert scale (like to dislike), ordering, navigational (receiving immediate feedback after changing preference for one item), and affective rating.

In summary, the mitigation approaches of confirmation bias and homophily use the visual display of information to increase users' awareness for item-features of interest (e.g., position valence, similarity,..) and to encourage and facilitate the intuitive exploration of diverse items. Approaches include multidimensional feature representation plots, and additional highlighting in form of color-coding or clustering of meaningful feature combinations. Two studies aim to enable users to understand contingencies between preferences, item selections and recommendation outcome and thus to a certain degree explaining recommendations. They do this by visually displaying the system's user model in form of a word cloud or an interest profile, preference summary, value chart or outcome view.

(RQ4) *Evaluation*: On their attempt to mitigate confirmation bias, Liao and Fu (2014) measured the potentially moderating factor of accuracy motive (motivation to accurately learn about a subject) of the users before exposure to the online forum. Results of the user study show that accuracy motive and position magnitude (moderate/extreme) of authors were functioning as moderating factors by influencing the effectiveness of bias mitigation. The authors conclude that interfaces should be individually adapted for users with varying levels of accuracy motive and that authors with moderate opinion could function as bridges between users with different opinions. Graells-Garrido et al. (2016)'s clustered visualization of recommendations, aiming to mitigate confirmation bias and homophily, was found to be effective in increasing users' exploration behavior (users clicked on more diverse items). The proposed recommendation algorithm based on shared latent topics, however, was not effective in increasing exploration behavior. The results show that political involvement of the users was functioning as a moderating factor, influencing the effectiveness of bias mitigation. Thus, Graells-Garrido et al. (2016) conclude that no one-size-fits-all solution exists, but that indirect approaches of transparent recommendations and user profiles rather than directly exposing users to opposing information should be considered for bias mitigation. Results of Tsai and Brusilovsky (2017)'s study on mitigating homophily and position biases show, that the exploration patterns were more diverse in the experimental conditions of presenting scientists in a multi-dimensional scatterplot compared to a baseline of displaying them in a ranked list. However, in a post-experimental questionnaire users reported a higher intent to reuse the ranked list than the multi-dimensional scatterplot. The authors conclude that diversity-oriented interfaces on the one hand can encourage the exploration of more diverse recommendations, but on the other hand can also impair intent to reuse the system and thus should be designed with care. The results of Pommeranz et al. (2012)'s user study on mitigating framing, anchoring and loss aversion during preference elicitation, show cognitively less demanding rating tasks were liked most and resulted in highest quality outcome lists. They conclude, that the interface design needs to adapt to individual differences in terms of user preferences. The authors highlighted the importance of transparency and control on the grounds that users found it very useful to be allowed to investigate the links between their interests, preferences and recommendation outcomes.

In summary, multiple studies highlight that no one-size-fits-all mitigation approach exists due to moderating user-related factors, such as the accuracy motive, diversity seeking or challenge averseness, motivation, political involvement and opinion. Thus the authors emphasize that interfaces should thus be designed to be personalizable. In addition, the need for transparent and interactive interface designs which allow control of user-profile and recommendations was highlighted.

4 Discussion

In this paper, we reviewed interface-based approaches for the mitigation of confirmation bias, homophily, position bias, framing, anchoring, and loss aversion **(RQ1)**. To measure bias, the studies compared the effect of an intervention with a baseline system on a set of metrics **(RQ2)**. The reviewed studies applied interactive multidimensional visualizations, rearranging, sorting, and highlighting through color-coding and size to increase users' awareness for diverse features, to facilitate and increase exploration of recommended items, and to align the system's user model with the user's mental preference model **(RQ3)**. During the evaluation of the approaches **(RQ4)**, multiple user-related factors that *moderated* the effectiveness of the reviewed mitigation approaches were identified. Consequently, the studies highlighted the need for personalized interfaces that can adapt to these factors. They include users' accuracy motive, motivation, political involvement, and prior opinions on recommended items or topics, all measured with tailor-made questionnaires or inferred from the user's behavior. Overall, transparency, control, as well as immediate feedback were found to enhance the users' understanding and to mitigate cognitive bias.

While the surveyed methods are within graphical interfaces, they help to uncover research questions for future studies in all interactive interfaces, also for *natural language-based* mitigation strategies:

1. Which approaches of interactive natural language bias mitigation approaches are most effective?
2. In which form and to which extent should transparency and control be given to the users?
3. What are user-related moderating factors and how could they be measured?

4. How could an interface personalization according to these user-related factors look like?

Our literature review also suggests that bias mitigation strategies using natural language could be used at different stages of interaction: **a)** conversational preference elicitation, **b)** pre-evaluation and explanation of recommended items, or **c)** to motivate behavior modifications for bias mitigation. Such interactions could promote the users' understanding of their profiles and the functioning of the system. Using NLG to increase user-control on the user-profile, algorithmic parameters, and the recommendation outcomes (Jin et al., 2020), appears to be a promising way to mitigate cognitive biases.

5 Conclusion

The analysed studies demonstrate effective approaches of implementing and evaluating interface-based cognitive bias mitigation for recommender system users. On this basis, we suggest promising areas for future research for bias mitigation using interactive NLG: personalization of explanations, and more immediate transparency and control.

Acknowledgments

This work has received funding from the *European Union's Horizon 2020* research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860621.

References

Ricardo Baeza-Yates. 2018. Bias on the web. *Communications of the ACM*, 61(6):54–61.

Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132.

Engin Bozdag and Jeroen van den Hoven. 2015. Breaking the filter bubble: democracy and design. *Ethics and Information Technology*, 17(4):249–265.

Mauro Dragoni, Ivan Donadello, and Claudio Eccher. 2020. Explainable ai meets persuasiveness: Translating reasoning results into behavioral change advice. *Artificial Intelligence in Medicine*, page 101840.

Eduardo Graells-Garrido, Mounia Lalmas, and Ricardo Baeza-Yates. 2016. Data portraits and intermediary topics: Encouraging exploration of politically diverse profiles. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 228–240.

Marco Guerini, Oliviero Stock, Massimo Zancanaro, Daniel J O'Keefe, Irene Mazzotta, Fiorella de Rosis, Isabella Poggi, Meiyii Y Lim, and Ruth Aylett. 2011. Approaches to verbal persuasion in intelligent user interfaces. In *Emotion-Oriented Systems*, pages 559–584. Springer.

Thomas T Hills. 2019. The dark side of information proliferation. *Perspectives on Psychological Science*, 14(3):323–330.

Yucheng Jin, Nava Tintarev, Nyi Nyi Htun, and Katrien Verbert. 2020. Effects of personal characteristics in control-oriented user interfaces for music recommender systems. *User Modeling and User-Adapted Interaction*, 30(2):199–249.

Q Vera Liao and Wai-Tat Fu. 2014. Can you hear me now? Mitigating the echo chamber effect by source position indicators. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 184–196.

Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Alina Pommeranz, Joost Broekens, Pascal Wiggers, Willem-Paul Brinkman, and Catholijn M Jonker. 2012. Designing interfaces for explicit preference elicitation: a user-centered investigation of preference representation and elicitation process. *User Modeling and User-Adapted Interaction*, 22(4-5):357–397.

Chun-Hua Tsai and Peter Brusilovsky. 2017. Leveraging interfaces to improve recommendation diversity. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 65–70.