Interactive Interventions to Mitigate Cognitive Bias

Rieger, Alisa

DOI
10.1145/3503252.3534362

Publication date
2022

Document Version
Final published version

Published in
UMAP2022 - Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization

Citation (APA)

Important note
To cite this publication, please use the final published version (if applicable). Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights. We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.
Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' - Taverne project

https://www.openaccess.nl/en/you-share-we-take-care

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.
Interactive Interventions to Mitigate Cognitive Bias

Alisa Rieger
a.rieger@tudelft.nl
Delft University of Technology
Delft, Netherlands

ABSTRACT

While the web offers a great potential to find and share information, the cognitively demanding conditions of online interactions can leave users vulnerable to cognitive biases, such as the confirmation bias – the tendency to favor information that confirms prior attitudes and beliefs when searching for, selecting, interpreting, sharing, and recalling information. This can negatively impact individuals’ decision-making and is likely to drive ideological polarization and extremism. With my dissertation, I am investigating whether and how interactive bias mitigation interventions, with a special focus on confirmation bias, could empower web users in making informed, unbiased, and autonomous choices. Based on my findings and observations, I plan to build a framework of user- and context-adaptive bias mitigation approaches during different kinds of web interactions.

CCS CONCEPTS
• Human-centered computing → User studies; • Information systems → Personalization.

ACM Reference Format:

1 SUPERVISION AND ORGANIZATION

I am Alisa Rieger, a PhD Candidate at Delft University of Technology (enrolled in TU Delft Graduate School) and am supervised by Professor Nava Tintarev from Maastricht University and Dr. Mariët Theune from University of Twente (both in the Netherlands). I am pursuing a PhD with a four-year program in full-time, of which I currently in the second year (projected completion date: 01.06.2024). Further, I am a Marie Skłodowska Curie Early Stage Researcher within the NL4XAI (Natural Language for Explainable Artificial Intelligence) project.

2 MOTIVATION AND CONTEXT

The web offers a great potential to find and share information, deliberate with fellow users, and educate oneself. The information we find online impacts trivial everyday decisions, as well as important life decisions [5]. However, the open nature of the web also comes with downsides: low quality and non-expert content including dis- and misinformation and overwhelming amounts of possible choices of what content to interact with. These conditions, combined with the limited amount of time users have for, or are willing to spend with, web interactions (i.e., web search, online debates, social media), are causing a high cognitive demand. This information overload leaves users vulnerable to cognitive biases during the interaction behavior itself, and the subsequent decision making and attitude forming [1, 8, 11, 18, 19, 25]. One type of bias that stands out as being specifically relevant across web applications is the confirmation bias, the human tendency to favor information that confirms prior attitudes and beliefs when searching for, selecting, interpreting, sharing, and recalling information [17]. Cognitive biases such as the confirmation bias can negatively impact individuals’ decision making [25], and on a societal level, are likely to drive ideological polarization and extremism [11, 14]. Cognitive biases that affect users’ web interactions can have an additional negative impact by causing second-order algorithmic biases which fuel the vicious cycle of bias in the web [2]. Such second-order biases are linked to recommender and ranking algorithms (i.e., for content recommendations and search engines) which filter and rank items according to their relevance to support the user in navigating the web. To improve recommendations and ranking, developers attempt to increase user interaction (i.e., mouse clicks, time spent), thus cognitive biases that affect user interaction would get amplified [2]. Further, platform operators do not necessarily attempt to improve the relevance of items to users’ intentions and promote autonomous choice, but to capture user attention to increase revenues, raising serious ethical questions [15]. Consequently, the problem of cognitive biases during web interactions has far-reaching negative implications for individuals and society and concerns multiple stakeholders (i.e., users, content providers, platform operators, society, legislators) with deviating interests.

To give back power to the web user for informed, unbiased, and autonomous choice, Lorenz-Spreen et al. [15] propose effective web governance by applying behavioral interventions. Since users’ web interaction behavior, their susceptibility to cognitive biases, and the reaction to cognitive bias mitigation approaches is known to be affected by situational (i.e., knowledge, interest) and stable (i.e., Need for Cognition) user-related factors [12, 13, 21, 22], such interventions likely need to be personalized by adapting to these factors to be effective for all users. With my dissertation I plan to investigate how this proposal could be implemented and to make the following contributions: Novel insights on confirmation bias mitigation for different users during web interactions derived from user studies and literature reviews. These insights build the basis for developing a framework of user- and context-adaptive and user autonomy preserving confirmation bias mitigation approaches.
for human computer interaction researchers and interface design guidelines to propose to practitioners.

2.1 Related work on Cognitive Bias Mitigation

To empower users to make well-informed choices and decisions during web interactions, Lorenz-Spreen et al. [15] propose effective web governance through the application of behavioral interventions in form of nudging or boosting.

Nudging and Boosting. Nudging refers to the approach of subtly modifying the choice architecture to alter people’s behavior towards optimal decision making, i.e., by hiding or highlighting certain options of the interface [23]. Caraban et al. [4] categorized different nudging approaches that aim at cognitive bias mitigation according to their transparency (non-transparent, transparent) and mode of thinking engaged (automatic mind, reflective mind), following Hansen and Jespersen [9]. While nudging, especially with approaches that are non-transparent and tap into the automatic mind, is criticized for the risk of paternalism, manipulation, and the lack of learning, boosting attempts to overcome these downsides by attempting to teach and empower users to become resistant to various pitfalls of web interactions, including cognitive biases [10, 15]. However, given that susceptibility to cognitive biases is increased by high cognitive demand, nudging approaches that prompt reflective choice and boosting approaches might further increase the cognitive demand and thus not always be the most suitable approach to mitigate biases. An approach that permits less cognitive demanding automatic and non-transparent nudges while sidestepping the risk of paternalism and manipulation was very recently proposed by Rejula and Hertwig [20]: Self-nudging through interventions that empower users to better reach their goals by designing their own choice-architecture.

User-related Factors and Personalization. Web interaction behavior, the susceptibility to cognitive biases, and the reaction to nudging approaches is affected by situational and stable user-related factors. A highly relevant stable factor, closely related to the notion of cognitive style, is the Need for Cognition (NFC), described as the individual’s tendency to organize their experience meaningfully [6]. NFC affects how users interact with information, to which extent this behavior is affected by cognitive biases, how they process explanations, and how they react to nudging and boosting approaches [3, 16, 24]. Bias mitigation can not be achieved with one-size-fits-all approaches, but likely requires to be adapted to the aforementioned user-related factors [21].

3 RESEARCH OBJECTIVES

My main research objective is to identify approaches that effectively mitigate cognitive biases during opinion-forming and decision-making processes, when selecting, interpreting, and drawing conclusions based on information items during web interactions for all users without harming their autonomy. To achieve this objective, I plan to investigate natural language based interactive nudging and boosting interventions (i.e., explanations, warning labels, summaries; see Figures 1 and 2) that target a combination of automatic and reflective thinking to effectively mitigate bias, with a primary focus on confirmation bias, during web interactions (i.e., web search, online debates).

Effective Mitigation and User Autonomy. Nudging approaches have been criticized for the risk of paternalism, manipulation, and the lack of learning. Thus, I want to investigated how interactive nudging, self-nudging, and boosting approaches can be applied and combined to effectively mitigate users’ cognitive biases while avoiding a decrease in user autonomy. (RQ1) How should reflective and automatic mitigation approaches be combined to effectively mitigate bias without decreasing user autonomy?

Personalization. Web interaction behavior and the susceptibility to cognitive biases is affected by situational and stable user-related factors. Consequently, mitigation approaches should be adapted to such user-related factors to be effective for all users. To achieve this, I aim at investigating the following research question: (RQ2) What user-related factors affect susceptibility to cognitive bias and the effectiveness of different mitigation approaches in which way? How should mitigation approaches be personalized to be efficient for all users?

Metacognition and Learning. Successful bias mitigation is mostly defined solely based on observed user behavior. However, I argue that additional measures should be considered, such as users’ accurate self-assessment and awareness of bias in their behavior (metacognition), or whether (long-term) learning is facilitated by an approach. (RQ3) What constitutes effective bias mitigation, including measures beyond users’ direct interaction behavior, i.e., their awareness of their own bias, or their (long-term) learning?

Bias Predictors. Susceptibility to cognitive biases is dependent on context-related factors and the cognitive demand they cause for different users (RQ2), such as the elements of the user-interface, the nature and intended use of the application, or the quantity and quality of contributions from fellow users. Learning which conditions indicate an increased risk of bias for which user is highly valuable to establish in which situations bias mitigation approaches should be applied, and what aspects of the web interaction they should attempt to enhance or counter. (RQ4) What conditions indicate an increased risk of cognitive bias and thus require and should be targeted by mitigation approaches?

Research-Methods. I investigate the above research questions primarily by means of literature reviews and user studies. For this, online studies (as compared to lab studies) are particularly useful, since they permit the collection of declarative data (questionnaires and behavioral data (i.e., mouse movements, time spent, items clicked) with mock web applications from many participants with diverse backgrounds (i.e. age, country of origin, level of education). The material required for the user studies (i.e., viewpoint labels for search results) are collected with crowd computing. For all studies, I follow open science principles and pre-register hypotheses, the research-, and analysis plan prior to collecting data.

4 RESULTS TO DATE

So far, I have conducted a literature review and two extensive user studies. The literature review was published in a workshop and investigated prior work on cognitive bias mitigation during web interactions [21], insights of which on how to measure bias, and how to...
evaluate whether an approach is effective, informed the design of the user studies. The first study investigated the effect of obfuscations (i.e., hiding the result unless the user clicks on it) with warning labels on interaction with attitude-confirming search results (see Figure 1). With this obfuscation study, we explored whether participants’ cognitive style might have a moderating or mediating effect on the intervention and affect users’ awareness of bias in their interaction behavior. The second user study (summary study) (under review) investigated the effect of personalized persuasive suggestions to motivate users to engage with debate summaries. The influence of these persuasive strategies was evaluated in terms of participants’ recall of attitude-opposing arguments made in the debate (see Figure 2). In the following, I present the observations we made so far in light of the research questions.

(RQ1) Effective Mitigation and User Autonomy: In the obfuscation study we found that targeted obfuscations decrease interaction with attitude-confirming search results. We further observed that, even when applied to random search results, obfuscations reduced interaction with those search results and reason that this was likely due to participants ignoring obfuscated items due to the decreased ease of access (automatic), which would indicate a harm in user autonomy for item selection. In the summary study, we did not observe confirmation bias across conditions and thus could not observe any effect of summaries and suggestions. (RQ2) Personalization: Our exploratory observations in the obfuscation study suggest that cognitive style might have an effect on engagement with the warning labels. They further indicate that, depending on the cognitive style of a user, the effectiveness and cognitive mechanism (reflective/automatic) prompted by the approach deviates. In the summary study we investigated personalized persuasive suggestions to engage with the summary. However, due to not observing confirmation bias across conditions, we could not find an effect of bias mitigation. (RQ3) Metacognition and Learning: In the obfuscation study, we observed that participants who saw a warning label overestimated and likely overcorrected the bias in their behavior more than those who did not see a warning label. This observation first prompted RQ3, namely, to consider measures beyond the mere interaction behavior when determining whether bias mitigation approaches are effective (also see [7]). We followed this consideration when designing the summary study, in which we measured participants’ argument recall as a measure of what items participants engaged with on a level deep enough that it could build the foundation for long-term learning. (RQ4) Bias Predictors: This research question is inspired by the high recall found in the summary study, regardless of persuasive and mitigation strategy. These findings raised the question whether the experimental set-up created a context with a low risk for cognitive bias. We argue that we might not have observed confirmation bias due to participants’ high investment in the task, the high quality of contributions, and the absence of distractors.
5 EXPECTED NEXT STEPS AND FUTURE DIRECTIONS

With the next concrete steps I plan to get insights into user- and context-adaptation of bias mitigation approaches. First, I am preparing a follow-up to the obfuscation study, motivated by the exploratory observations on the relation between cognitive style and the effect of automatic compared to reflective aspects of the warning label and obfuscation (RQ2). Second, I plan to investigate potential bias predictors (RQ4), since understanding the factors that indicate the level of bias risk enables us to apply bias mitigation approaches more effectively. For this user study, I am planning to implement a dialogue setting (online chat) between two participants who have to exchange information given to them in order to find the correct response to a task by following a set of rules. The response can be either correct or incorrect, permitting easy bias detection (confirmation bias – do participants stick to their initial (incorrect) opinion, and automation bias – do participants adopt the system’s (incorrect) opinion). This setting further permits simple manipulations of the experimental conditions, such as what information (correctness, alignment, completeness, granularity) will be given (system, colleague, superior, inferior) and when (once, successive) to whom of the participants. I am specifically interested in potential bias predictors that might be observed in participants dialogue, such as linguistic misalignment (missing adjustment of language between conversational partners). Findings from this study can inform approaches for other web applications in which users interact with fellow users (i.e., online debates, social media comments). Below, I describe how I plan to approach the research questions.

(RQ1) Effective Mitigation and User Autonomy: I plan to investigate the potential design space of concrete (self-) nudging and boosting approaches that might effectively mitigate cognitive bias during web interactions while preserving user autonomy. The expected outcome is an overview of the design space for confirmation bias mitigation approaches during web interactions that builds the basis for the framework of user- and context- adaptive bias mitigation. (RQ2) Personalization: For this RQ, I will investigate which specific automatic/reflective combined nudges are effective and in which way they affect user autonomy for users with different cognitive style. The expected outcome for this RQ will consist of guidelines on how to combine automatic and reflective aspects of a nudge for effective bias mitigation for different users. This will supply the part of user-adaptation to the framework of user- and context- adaptive bias mitigation. (RQ3) Metacognition and Learning: This research question has emerged from the observations made so far, however has not been investigated yet. To approach it, I plan to review literature on learning, and meta-cognition (i.e., self-assessment, awareness) with regards to decision making and opinion forming while and after interacting with information items. Further, I plan to explore users’ awareness of bias and learning during and after a web interaction with pre- and post-interaction questionnaires when investigating bias mitigation approaches in user studies. An expected outcome is a novel definition of effectiveness when evaluating bias mitigation approaches by including measures beyond directly observable behavior. (RQ4) Bias Predictors: I will investigate this research question in a first step with a user study I am currently preparing that focuses on potential bias predictors in metrics describing user dialogue, such as lexical misalignment. Further, I plan to approach this question by exploring what context-related factors, such as content of information items, interface elements, or mode of user- and system interactions (i.e., warning messages in natural language delivered by the system). An expected outcome is a set of guidelines indicating under which conditions which specific factors should be targeted by bias mitigation approaches. This will supply the part of context-adaptation to the framework of user- and context- adaptive bias mitigation.

6 DISSERTATION STATUS AND LONG-TERM RESEARCH GOALS

With the planned contributions of my dissertation, I hope to advance the attempt of empowering and supporting web users for an informed, unbiased, and autonomous decision making process during and after engaging with information items. For an overview of the current status of my dissertation, see Table 1. Next to answering the research questions by investigating the effectiveness
of potential mitigation approaches, I plan to consider ethical implications of the proposed approaches. These include, amongst others, whether an approach harms user autonomy, is vulnerable to misuse with malicious intentions, jeopardizes data protection rights, or elicits alternative cognitive biases (i.e., automation bias). From a perspective of long-term societal impact, I aspire that my contributions add to a body of research that motivates regulatory attempts, aiming at promoting the potential of the web for activities such as information retrieval and sharing, community building, or deliberation, and guiding platform providers to prioritize societal good over profit.

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

REFERENCES