This Item Might Reinforce Your Opinion
Obfuscation and Labeling of Search Results to Mitigate Confirmation Bias

Rieger, Alisa; Draws, Tim; Theune, Mariët; Tintarev, Nava

DOI
10.1145/3465336.3475101

Publication date
2021

Document Version
Final published version

Published in
HT 2021 - Proceedings of the 32nd ACM Conference on Hypertext and Social Media

Citation (APA)
Rieger, A., Draws, T., Theune, M., & Tintarev, N. (2021). This Item Might Reinforce Your Opinion: Obfuscation and Labeling of Search Results to Mitigate Confirmation Bias. In *HT 2021 - Proceedings of the 32nd ACM Conference on Hypertext and Social Media* (pp. 189-199). (HT 2021 - Proceedings of the 32nd ACM Conference on Hypertext and Social Media). ACM. https://doi.org/10.1145/3465336.3475101

This Item Might Reinforce Your Opinion

Obfuscation and Labeling of Search Results to Mitigate Confirmation Bias

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

Tim Draws
t.a.draws@tudelft.nl
Delft University of Technology
Delft, Netherlands

Mariët Theune
m.theune@utwente.nl
University of Twente
Enschede, Netherlands

Nava Tintarev
n.tintarev@maastrichtuniversity.nl
Maastricht University
Maastricht, Netherlands

ABSTRACT

During online information search, users tend to select search results that confirm previous beliefs and ignore competing possibilities. This systematic pattern in human behavior is known as *confirmation bias*. In this paper, we study the effect of *obfuscation* (i.e., hiding the result unless the user clicks on it) with warning labels and the effect of *task* on interaction with attitude-confirming search results. We conducted a preregistered, between-subjects crowdsourced user study (N=328) comparing six groups: three levels of obfuscation (targeted, random, none) and two levels of task (joint, two separate) for four debated topics. We found that *both* types of obfuscation influence user interactions, and in particular that targeted obfuscation helps decrease interaction with attitude-confirming search results. Future work is needed to understand how much of the observed effect is due to the strong influence of obfuscation, versus the warning label or the task design. We discuss design guidelines concerning system goals such as decreasing consumption of attitude-confirming search results, versus nudging users toward a more analytical mode of information processing. We also discuss implications for future work, such as the effects of interventions for confirmation bias mitigation over repeated exposure. We conclude with a strong word of caution: measures such as obfuscations should only be used for the benefit of the user, e.g., when they explicitly consent to mitigating their own biases.

CCS CONCEPTS

• **Human-centered computing → User studies**; • **Information systems → Search interfaces**.

KEYWORDS

Confirmation Bias, Web Search, Warning Labels, Obfuscation, Nudging, Cognitive Bias Mitigation

HT '21, August 30–September 2, 2021, Virtual Event, Ireland.
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8551-0/21/08…$15.00
https://doi.org/10.1145/3465336.3475101

ACM Reference Format:
Alisa Rieger, Tim Draws, Mariët Theune, and Nava Tintarev. 2021. This Item Might Reinforce Your Opinion: Obfuscation and Labeling of Search Results to Mitigate Confirmation Bias. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media (HT '21), August 30–September 2, 2021, Virtual Event, Ireland.* ACM, New York, NY, USA, 11 pages. https: //doi.org/10.1145/3465336.3475101

1 INTRODUCTION

Previous research has shown that, during online information search, users tend to select search results that confirm pre-existing beliefs or values and ignore competing possibilities [5] (i.e., a systematic pattern in human behavior known as the *confirmation bias* [40]). This behavior, however, is likely to increase ideological polarization and extremism [17, 33]. Susceptibility to biases such as the confirmation bias has been linked to a lack of analytical thinking, as has susceptibility to misinformation [42]. Given this parallel, our approach to confirmation bias mitigation is inspired by efforts to mitigate the spread of misinformation: research showed that displaying warning labels prior to exposure to misinformation and requiring users' active consent before showing the item is effective in stimulating more skepticism, analytic information processing, and decreasing the interaction with misinformation [22, 26, 31, 34].

We investigated whether showing warning labels prior to exposure to attitude-confirming search results (i.e., search results which support a viewpoint in line with a user's attitude on a topic) could be likewise effective in mitigating confirmation bias during online search. We thus aimed to achieve a decrease in confirmation bias during search by applying search result obfuscations with warning labels. This way, we wanted users to look at a topic from different viewpoints and, consequently, make more informed decisions.

This research addresses the following question: *Can search result obfuscations with warnings of confirmation bias mitigate confirmation bias by motivating users to interact with attitude-opposing search results during search for information on debated topics?* We investigated this question by conducting a preregistered, between-subjects user study with crowd-workers.[1] In this study, we observed user interaction (N=328) with search results on four different debated

[1] Preregistering meant publicly determining our hypotheses, experimental setup, and analysis plan before any data collection. The (time-stamped) preregistration document can be found in our repository: https://osf.io/32wym/.

topics, comparing the interaction behavior between six groups (three levels of search result display, and two levels of task).

Our results show that obfuscating search results with warning labels is effective in decreasing interaction with these search results. We also found that targeted obfuscations of attitude-confirming search results causes increased interactions with attitude-opposing search results and thus might be an effective approach of mitigating confirmation bias during search result selection. In sum, we make the following contributions:

(1) **Viewpoint annotated search results:** we provide a data set with 200 search results on four topics which are viewpoint annotated on a seven-point Likert scale ranging from "strongly opposing" to "strongly supporting"
(2) **Preregistered user study**: we conduct a preregistered between-subject user study (n = 328) to investigate the effect of obfuscations with warning labels on the interaction with search results on debated topics
(3) **Rich data set of interaction behavior**: we provide a data set with interaction behavior (clicks and markings) of 328 participants with search results on four disputed topics for three conditions of search result display and two levels of task design
(4) **Ethical considerations**: we discuss ethical considerations of the approach for confirmation bias mitigation during search that we investigated with this study (motivated by our findings)
(5) **Design guidelines:** we propose design guidelines for confirmation bias mitigation during search with targeted search result obfuscations with warning labels in light of our findings and ethical considerations

The preregistration, data-sets, and material for gathering the data as well as for analyzing the results and replicating our study are publicly available.[2]

2 RELATED WORK AND HYPOTHESES

In this section, we look at findings on confirmation bias during search and cognitive bias mitigation. Further, we take a look at approaches to nudge users to a more analytic information processing which were applied to combat online misinformation and at potential user-related factors that might result in behavioral patterns during search. Note that the hypotheses we present here had been preregistered before any collection of data.

2.1 Confirmation Bias During Search

Online search for information has become an indispensable part of our day-to-day life, whether we are trying to settle trivial discussions, looking for information on how best to do something, or collecting information before making the decision on who to vote for in an election. Online information thus affects our decisions, even important life decisions, immensely [7]. To make an efficient decision despite the overwhelming amount of possible choices of search results, we tend to apply search strategies, for example by searching for information which confirms our prior beliefs [5, 28]. Even though these strategies help us in many cases to make faster and easier decisions by reducing the amount of information and uncertainty [14], they can also do harm when we have the intent to make a well-informed decision but miss out on information supporting another viewpoint than our own. In a broader perspective, such behavior is likely to drive polarization, diminish the quality of public discourse, and contribute to ideological extremism [17, 33]. Regardless of the specific democratic theory one supports, nearly all strands of democratic theory emphasize the importance of promoting viewpoint diversity [39].

When identifying how to mitigate confirmation bias during search, the search process can be divided into three sub-processes during which confirmation bias can occur in different forms: (1) querying for, (2) selecting, and (3) making decisions based on information [24]. We focus on the process of (2) selecting information, during which confirmation bias can be observed in an increased likelihood of interaction (i.e., clicking on or sharing) with search results that confirm our prior beliefs compared to competing possibilities [2, 5]. A widely used measure of confirmation bias during search result selection is thus the number/proportion of selected attitude-confirming search results compared to attitude-opposing ones [24, 28]. In this study, we investigate information selection on two levels (see Section 2.3): for oneself by clicking on items (i.e., *clicking behavior*) and for others by sharing items (i.e., *marking behavior*).

2.2 Nudges for Confirmation Bias Mitigation

The concept of nudging refers to mechanisms that subtly influence users to make decisions which are considered to be beneficial for them, without restricting possible choices [52]. For confirmation bias mitigation during search, nudges could be applied in an indirect approach aiming at generally motivating analytical thinking and supporting users in being more susceptible to genuine evidence, referred to as *nudges for reason* by [30], and in a more direct way aiming at influencing users' item selection behavior and guiding them towards interaction with attitude-opposing search results. This can be achieved by applying nudges which aim at modifying the *Decision Structure*; e.g., by ordering items, setting defaults, or altering the required effort [21].

Prior work on confirmation bias mitigation mostly researched approaches of nudging towards a less biased item selection by means of data visualization [15, 32]. Nudges for bias mitigation based on natural language which may generate more immediate transparency for the users [49], have not been studied for confirmation bias mitigation yet. Such nudges have, however, been applied to guide users towards item selection or avoidance for the purpose of combating online misinformation. Previous work on this subject applied warning labels to flag items which may contain misinformation and decreased the ease of access by obfuscating these items by default [26, 31]. This way, users are effectively and transparently nudged towards increased scepticism of and decreased engagement with misinformation [8, 34]. Kaiser et al. [22] found that engagement was further decreased when requiring additional effort such as actively clicking a button. Investigating a similar approach in a context of bias mitigation, Hube et al. [20] found that presenting messages which explicitly make workers aware of potential bias and require interaction to proceed with the task are

[2] See link in Footnote 1.

effective in mitigating bias stemming from worker opinions during crowd-workers labeling tasks. We thus expected that this approach would be likewise effective in mitigating confirmation bias during information selection for oneself (*clicking*) and for others (*marking*) and formulated the following hypotheses:

H1: Users of search engines are less likely to *click* on attitude-confirming search results when some search results on the search engine result page (SERP) are obfuscated with a warning label.

H2: Users of search engines are less likely to *mark* attitude-confirming search results as particularly relevant when some search results are displayed with a warning label.

Next to the search result display, the intention users have when selecting search results is likely to affect their interaction. We will proceed to discuss relevant literature on task design for bias mitigation in the next section.

2.3 Effects of Task Design

Cognitive biases are likely to decrease or disappear if task or context stimulate more analytic information processing, for example by triggering high personal accountability or critical thinking in the user [18, 37, 51]. Further, user-studies testing a new interface feature such as the one we are presenting here might result in increased interaction with novel features caused by participants' curiosity. This is undesired for this study, but has been taken advantage of for nudging users towards certain actions in other studies [19, 53]. Another factor impacting the effectiveness of obfuscations with warning labels is repeated exposure to them. This might lead to initial strengthening, and over a longer term to habituation and thus weakening of their effect on users' interaction with attitude-confirming search results [4, 48].

To detect potential undesired effects of task design, curiosity, or repeated exposure to the warning label we asked participants to complete two sub-tasks, either in *two separate tasks* or in *one joint task*: (1) explore the SERP as they would do normally and, as a basis of sharing, (2) mark results they considered to be particularly relevant (for a detailed reasoning for this task design see Section 3.3). This led us to the following hypothesis:

H3: In the two separate tasks condition, users of search engines are less likely to *mark* attitude-confirming search results as particularly relevant, compared to the joint task condition.

2.4 User-related patterns in search behavior

In addition to external factors discussed in the previous sections, search result selection can be driven by internal factors which can be both situational or stable and are individually different for different users. Situational factors include factors such as the attitude strength, attitude certainty, and the interest in the topic. Strong attitudes and high certainty were shown to result in increased confirmation bias [27] and high interest was shown to result in increased information processing capabilities and consequently in more effective information processing [50]. A stable internal factor driving search result selection is for example the extent to which users value diverse viewpoints or are challenge averse [2, 38]. Another stable internal factor influencing the reaction to the warning labels we propose for confirmation bias mitigation during search is the susceptibility to persuasive messages [3]. Both factors are closely related to the concept of *Need for Cognition* (NFC). NFC has been described as "The individual's tendency to organize his experience meaningfully" [9] and affects how users interact with information and to which extent this behavior is affected by confirmation bias, and how they process explanations and (persuasive) messages [6, 36, 54].

We thus anticipate that users will have a general propensity to interact with viewpoint confirming views, or sensitivity to confirmation bias and warning labels and formulate the following hypothesis:

H4: Users are likely to display a consistent pattern of behavior while *clicking* on and *marking* attitude-confirming search results (i.e. participants' marking behavior correlates with their clicking behavior).

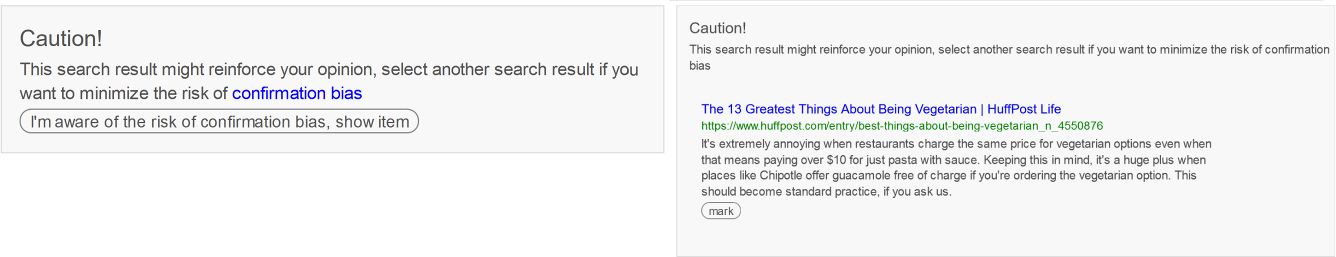
3 METHOD

To investigate our research questions outlined in Section 2, we conducted a between-subjects user study. We manipulated the factors **display** (*targeted obfuscation, random obfuscation, no obfuscation*) and **task** (*two separate tasks, joint task*) and evaluated the degree to which participants would click on and mark attitude-confirming search results.

3.1 Materials

3.1.1 Topics. Draws et al. [12] provide a data set containing user attitudes regarding 18 different controversial topics from the website *ProCon* [44]. These 18 topics were selected because the authors assumed that they would be applicable globally and that they would not include highly emotionally charged topics. The authors asked 100 participants to state their attitude towards each of these topics on a seven-point Likert scale ranging from "strongly disagree"(-3) to "strongly agree" (+3). From this data set, we selected topics for which we expected to observe confirmation bias; i.e., topics where participants reported to have comparatively large proportions of *strong* attitudes. We operationalized this as topics for which at least around 50% of participants selected the options -3, -2, +2, or +3. Following this criterion, four topics were included in the experiment: **(1)** Is Drinking Milk Healthy for Humans?; **(2)** Is Homework Beneficial?; **(3)** Should People Become Vegetarian?; **(4)** Should Students Have to Wear School Uniforms?

3.1.2 Search Results. Draws et al. [12] provided a data set of 50 search results for 14 pre-defined queries related to each of the topics using the *Bing API* [35]. From these 700 retrieved URLs per topic, we handpicked 50 opinionated search results by assessing their relevance for each of the four selected topics. The resulting 200 unique search results were subsequently annotated by crowd-workers on *Amazon Mechanical Turk* [55]. Specifically, workers annotated the relevance to the topic (binary) and the viewpoint with respect to the topic (scored on a seven-point Likert scale ranging from "strongly opposing" to "strongly supporting"). We collected three annotations for each search results and observed a satisfactory inter-annotator agreement (*Krippendorff's* α = 0.78) [16]. In our final data set, the search result was assigned the median value of these three annotations. Per topic, we subsequently selected 12 search results by randomly sampling two "strongly supporting", two "supporting", two "somewhat supporting", two "somewhat



Caution!

This search result might reinforce your opinion, select another search result if you want to minimize the risk of confirmation bias

Caution!

This search result might reinforce your opinion, select another search result if you want to minimize the risk of confirmation bias

The 13 Greatest Things About Being Vegetarian | HuffPost Life

https://www.huffpost.com/entry/best-things-about-being-vegetarian_n_4550876

It's extremely annoying when restaurants charge the same price for vegetarian options even when that means paying over $10 for just pasta with sauce. Keeping this in mind, it's a huge plus when places like Chipotle offer guacamole free of charge if you're ordering the vegetarian option. This should become standard practice, if you ask us.

Figure 1: Obfuscated search result during SERP exploration task before (left) and after (right) *show item* button was clicked.

opposing", two "opposing", and two "strongly opposing" from all search results that were deemed relevant by all crowd-workers.[3] They were displayed in random order (see Table 1).

3.1.3 Search Result Obfuscation. Search results were obfuscated with a warning label, warning of the risk of confirmation bias if this item is selected and advising the participant to select another item (see the left-hand panel in Figure 1). Here, the *Wikipedia* entry on confirmation bias [59] was linked so that participants could inform themselves about this cognitive bias. To view the obfuscated search result, participants had to click a button, stating they were aware of the risk of confirmation bias. After clicking the button, the search result would become visible underneath the warning label (see the right-hand panel in Figure 1). During the marking task in the *two separate tasks* condition, obfuscated search results were displayed in the same way (search result visible below warning label).

3.2 Variables

Independent variables

- *Display* (categorical, between-subjects). Participants were randomly assigned to one of three display conditions: (1) targeted obfuscation of moderate and extreme attitude-confirming search results (see Figure 1), (2) obfuscation of four randomly chosen search results, and (3) no obfuscation (see Table 1).
- *Task* (categorical, between-subjects). Participants were also randomly assigned to one of the two task conditions: (1) a *two separate tasks condition*, where search result exploration and marking particularly relevant results was split in two separate tasks or (2) a single *joint task condition*, where search result exploration and marking particularly relevant results were done together (see Figure 2).

Dependent variables

- *Click proportion attitude-confirming results* (continuous). Proportion of attitude-confirming results among the search results participants clicked on during search results exploration: [0,1].
- *Marking proportion attitude-confirming results* (continuous). Proportion of attitude-confirming results among the search results participants marked when asked for items they would share: [0,1].

Exploratory variables

- *Click proportion obfuscated search results.* For targeted and random obfuscation condition: proportion of obfuscated results among the search results participants clicked on during search results exploration.
- *Marking proportion obfuscated search results.* For targeted and random obfuscation condition: proportion of obfuscated results among the search results participants marked when asked for items they would share.

Descriptive variables

- *Gender.* Participants could select between "female", "male", or "non-binary/other".
- *Age.* Participants were asked to enter their age using a numerical value.
- *Time spent on the task.* Time participants spent on the whole task, including prior- and post-interaction questions.
- *Time spent on SERP exploration.* Time participants spent on the SERP exploring the search results.
- *Number of clicks.* Number of search results participants clicked on to retrieve the linked document.
- *Number of markings.* Number of search results participants marked as being particularly relevant.

3.3 Procedure

We conducted this study on the online survey platform *Qualtrics* [46]. To control for data quality, we integrated four attention checks into the survey (two prior to the clicking and marking task and two during post-interaction questions), asking participants to give a specific pre-defined response. The procedure, approved by the ethics committee of our institution, consisted of four subsequent steps (see Figure 2):

Step 1: Pre-interaction. Participants were given a short introduction to the experiment and asked to answer demographic questions (gender, age). We then asked them to imagine the following scenario: *"You had a discussion with a relative or friend on a certain topic. The discussion made you curious about the topic and to inform yourself further you are conducting a web search on the topic."* Subsequently, we asked participants to state their attitude towards the four selected topics (see Section 3.1.1) on a seven-point Likert scale ranging from "strongly disagree" to "strongly agree". The responses "strongly disagree", "disagree", "agree", and "strongly agree" were considered to be *strong attitudes*.

[3] Data sets containing the 12 included search results per topic as well as all 200 annotated search results are publicly available at link in Footnote 1.

Table 1: Representation of the search result display. Each row represents a search result (twelve in total), with two results per viewpoint (-3 to +3), displayed in random order (example); Obfuscation illustrated with [[ ]]. From left to right the columns represent the three conditions (with two variants of targeted obfuscation). *No Obfuscation*: No search result is initially obfuscated. *Targeted Obfuscation*: Depending on reported attitude, participants saw either sup or opp (supporting/opposing obfuscation). Moderate (±2) and extreme (±3) attitude-confirming search results are initially obfuscated. *Random Obfuscation*: Four randomly selected search results are initially obfuscated.

No Obfuscation	Targeted Obfuscation sup	Targeted Obfuscation opp	Random Obfuscation
-1	-1	-1	[[-1]]
+2	[[+2]]	+2	[[+2]]
-3	-3	[[-3]]	-3
-2	-2	[[-2]]	-2
+3	[[+3]]	+3	+3
-1	-1	-1	-1
+2	[[+2]]	+2	[[+2]]
+1	+1	+1	+1
-2	-2	[[-2]]	-2
+3	[[+3]]	+3	+3
+1	+1	+1	+1
-3	-3	[[-3]]	[[-3]]

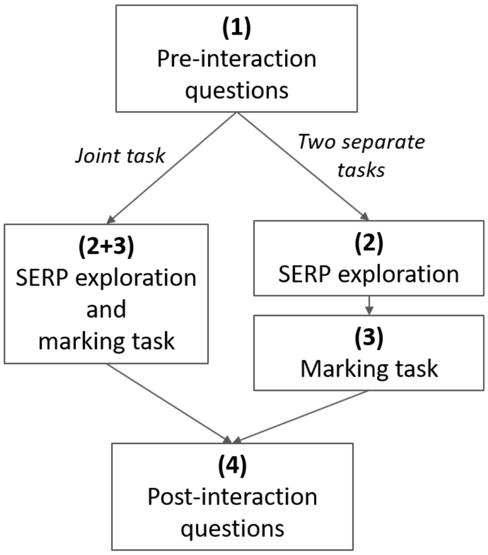

(1)
Pre-interaction questions
Joint task
Two separate tasks
(2+3)
SERP exploration and marking task
(2)
SERP exploration
(3)
Marking task
(4)
Post-interaction questions

Figure 2: Data collection procedure for *joint task* and *two separate tasks* condition: pre-interaction questions, SERP exploration, marking task, and post-interaction questions.

Step 2 and 3: Search result exploration and marking. Based on their answers during Step 1, participants were randomly assigned to (1) one of the topics they held a strong attitude towards,[4] (2) a **display** condition, and (3) a **task** condition. They were thus exposed to a randomly ordered list of search results relevant to their assigned topic in their assigned search result display format. Participants' task was to explore the search results (Step 2) and mark search results that they considered to be particularly relevant (Step 3). Depending on their assigned task condition, they would perform these actions separately or together:

- *Two separate tasks condition.* Participants saw the list of 12 search results (with obfuscations depending on their assigned **display** condition) relevant to their assigned topic (*SERP exploration*). They were given as much time as they wanted to explore the search results and examine the linked documents.[5] After continuing to the next page, participants were again presented with all 12 search results. Among those results, they were then asked to mark items that they considered to be particularly relevant and informative and that they would have liked to forward to a relative who wants to form an opinion on the topic (*marking task*). Search results which were obfuscated during SERP exploration were still displayed with the warning but not obfuscated (see Figure 1). Participants were not able to examine the linked documents again but could only see the titles and snippets.
- *Joint task condition.* As above, but participants could mark items that they considered to be particularly relevant and informative (*marking task*) at the same time as they explored the search results (*SERP exploration*).

Step 4: Post-interaction. We asked participants to state their attitude on the selected topic again and to answer a number of questions on their experience with the task, self-perception of their behavior, and user-experience.[6]

Reasoning for task design. To be able to draw valid conclusions from the collected data, we attempted to design a task and scenario that would motivate participants to mimic their natural search result exploration behavior by requiring a low level of accountability. However, since the feature of obfuscations with warning label was novel, we had to control for potential effects of curiosity. We did so by observing a second level of interaction behavior, for which we expected users to be less driven by curiosity because this second task required increased accountability. We thus observed two levels of participant behavior, first (1) exploring search results for themselves *(clicking: low accountability, potentially high curiosity)*, and then (2) for others *(marking: high accountability, low curiosity)*. However, by asking participants to mark particularly relevant search results in a separate task and displaying the warning labels again, we could have introduced an unwanted effect of repeated exposure to the warning labels. To allow us to single out potential

[4] Participants who did not hold a strong attitude towards any of the four topics were ejected from the study; see Section 3.4.

[5] We intentionally left the duration for exploration up to the participants to best mimic natural exploration behavior.

[6] Further details on the post-interaction questionnaire are available at link in Footnote 1.

undesired effects of task, curiosity, or repeated exposure to the warning label we asked participants to complete the two tasks, either in *two separate tasks* or in *one joint task*.

3.4 Sample

Before collecting data, an a priori power analysis for a between-subjects ANOVA (with $f = 0.25$, $\alpha = \frac{0.05}{4} = 0.0125$ (i.e., due to testing four different hypotheses), and (1- β) = 0.8) determined a required sample size of 282 participants.

We initially recruited a total of 510 participants via the online participant recruitment platform *Prolific* [45]. Participants were required to be at least 18 years old and to speak English fluently. They were allowed to participate only once and were paid £1.75 for their participation (*mean* = £7.21/h). From these 510 participants, 182 were excluded from data analysis because they did not fulfil the inclusion criteria: they did not report to have a strong attitude on any of the topics *(41)*, failed at one or more of four attention checks *(50)*, spent less than 60 seconds on the SERP (*80*), or did not click on and mark any search results *(11)*.

Of the remaining 328 participants (gender: 49% female, 51% male, <1% non-binary/other; age: $mean = 28.8, sd = 10.6$), 282 clicked on at least one search result and thus were included in testing **H1** (clicking behavior), 293 marked at least one search result and thus were included in testing **H2** (marking behavior) and **H3** (task difference marking behavior), and 248 clicked on *and* marked at least one search result and thus were included in testing **H4** (correlation clicking and marking behavior). Participants were randomly assigned to one of the topics for which they reported to strongly agree or disagree (-3, +3). If they did not report to strongly agree or disagree for any of the four topics, they were assigned to a topic for which they reported to agree or disagree (-2, +2). If participants did not report a strong attitude (-3, -2, +2, +3) on any of the four topics, they were not able to participate further but received partial payment (£0.50).

3.5 Statistical Analysis

To test our hypotheses we planned to apply a one-way ANOVA to compare the clicking behavior between the three **display** conditions, and a two-way ANOVA two compare the marking behavior between the three **display** and the two **task** conditions. The terms clicking and marking behavior refer to the proportion of clicks on and markings of attitude-confirming search results. However, *Shapiro-Wilk* tests revealed that our observations were not normally distributed. Hence, we applied *Kruskal-Wallis* tests for testing **H1** (clicking behavior), **H2** (marking behavior), and **H3** (task difference marking behavior). For pairwise post-hoc testing of differences between **display** conditions, we applied *Dunn* tests. To test **H4** (relation clicking and marking behavior), we conducted a *Spearman's* rank correlation analysis. For testing all four hypotheses, the significance threshold was set at $\alpha = \frac{0.05}{4} = 0.0125$, aiming at a type 1 error probability of $a = 0.05$ and applying Bonferroni correction to correct for multiple testing. For the post-hoc *Dunn* tests for **H1** and **H2** of differences between the three **display** conditions, the significance threshold was set to $\alpha = \frac{0.05}{3} = 0.0167$ each, due to testing 3 pairwise comparisons. All analyses was conducted in *R* [13, 23, 25, 41, 47, 57, 58].

4 RESULTS

In the following section, we will present the collected data by means of descriptive statistics and the results of hypotheses testing as specified in Section 3.5.

4.1 Descriptive Statistics

Participants' distribution over the 6 different conditions (three display and two task conditions) was comparable: 49 to 66 participants completed the task in each condition. The criterion for topic assignment resulted in 64, 94, 68, and 102 participants being assigned to the topics 1, 2, 3, and 4, respectively. The average time spent on the task was 17.3 minutes ($se = 0.5$) with no difference between conditions. The time spent on the SERP page was 4.8 minutes for the **joint task** condition ($se = 0.3$) and 4.1 minutes for the **two separate tasks** condition ($se = 0.3$) with no differences between **display** conditions. We recall that 80 participants were excluded from the study for spending fewer than 60 seconds on the SERP, but note that these durations are substantially higher than 60 seconds.

With regards to the level of interaction, the mean number of clicks during search result exploration was 3.02 for the **joint task** condition ($se = 0.2$) and 2.57 for the **two separate tasks** condition ($se = 0.15$) with no differences between **display** conditions. This reflects roughly 3/12, or 25% of the search results. The mean number of markings was 2.95 ($se = 0.11$, no difference between conditions). This degree of interaction is consistent with the qualitative feedback which suggests that participants understood the task well, and found it interesting and enjoyable.

4.2 Hypotheses Testing

H1 - Obfuscations with warning labels result in lower proportion of clicks on attitude-confirming search results. The results of a Kruskal-Wallis test for the click behavior show evidence for a moderate effect of search result **display** on the proportion of attitude-confirming clicks ($H(2) = 33.87, p < .001, \eta^2 = .11$). A pairwise post-hoc Dunn test shows that the proportion of clicks on attitude-confirming search results was significantly lower in **targeted obfuscation** ($mean = 0.34, se = 0.03$) compared to **random obfuscation** ($mean = 0.54, SE = 0.03$; $p < .001$) and **no obfuscation** ($mean = 0.58, SE = 0.03$; $p < .001$; see Figure 3). However, there was no difference in the clicking behavior between the **random obfuscation** and **no obfuscation** conditions, leaving our hypothesis only partially confirmed.

H2 - Obfuscations with warning labels result in lower proportion of markings of attitude-confirming search results. The results of a Kruskal-Wallis test for the marking behavior likewise show evidence for a moderate effect of the factor **display** on the proportion of attitude-confirming markings ($H(2) = 21.23, p < .001, \eta^2 = .07$). A pairwise post-hoc Dunn test shows that the proportion of markings of attitude-confirming search results was significantly lower in **targeted obfuscation** ($mean = 0.47, SE = 0.03$) compared to **random obfuscation** ($mean = 0.65, SE = 0.03$; $p < .001$) and **no obfuscation** ($mean = 0.66, SE = 0.03$; $p < .001$; see Figure 3). As was the case for the clicking behavior, there was no difference in the marking behavior between **random obfuscation** and **no obfuscation**.

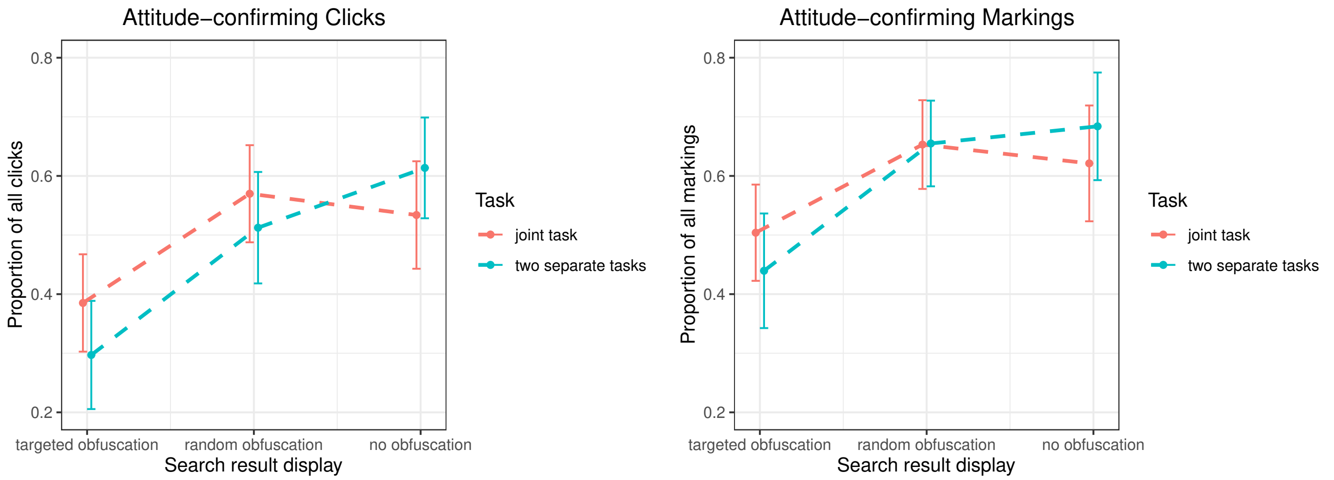


Figure 3: Interaction with attitude-confirming search results: mean proportion of participant's *(H1)* attitude-confirming clicks (left) and *(H2)* markings (right) per display condition (targeted obfuscation, random obfuscation no obfuscation) and per *(H3)* task condition (joint task and two separate tasks) with 95% confidence intervals. A proportion of one implies that all of a participant's clicks/markings were on attitude-confirming search results.

H3 - Two separate tasks condition results in lower proportion of markings of attitude-confirming search results. Against our hypothesis, the result of a Kruskal-Wallis test for the marking behavior does not show evidence for an effect of the factor **task** on the proportion of attitude-confirming search results ($H(2) = 0.04, p = .83$).

H4 - Behavioral pattern: Clicking and marking behavior are correlated. A Spearman rank correlation test shows evidence for a substantial positive correlation between the proportion of attitude-confirming clicks and markings ($\rho = .51, p < .001, R^2 = 0.26$). After controlling for the effect of **display** on the relationship, we still found clicking and marking behavior to be moderately positively correlated ($\rho = .44, p < .001, R^2 = 0.2$). This finding supports our hypothesis that participants would be displaying a consistent pattern of behavior across both tasks.

4.3 Exploratory results

Interaction with obfuscated search results. While obfuscated search results in the **targeted obfuscation** and the **random obfuscation** conditions make up a proportion of 33% of all displayed search results, the mean proportion of clicks on these search results is only 10% (see Figure 4). We observed no difference in this proportion of clicks between **targeted obfuscation** and **random obfuscation** conditions. For the marking behavior, this mean proportion is similar for the **joint task** condition ($mean = 0.12, SE = 0.02$), but higher for the **two separate tasks** condition ($mean = 0.21, SE = 0.03$).

5 DISCUSSION

In this study, we investigated whether item obfuscations with warning labels would be effective in countering confirmation bias in web search. We conducted a between-subject user study for which we manipulated the factors **display** *(targeted obfuscation, random obfuscation, no obfuscation)* and **task** *(two separate tasks, joint task).* We evaluated in which proportion the participants would click on and mark attitude-confirming search results.

Effect of display. While we found that targeted obfuscations with warning labels decreased the likelihood of interacting with (clicking, marking) attitude-confirming search results compared to no obfuscation, we did not find any effect of the random obfuscation condition. This finding implies that the mere presence of warning labels does not motivate users to decrease interaction with attitude-confirming search results, but that targeted obfuscations are required to achieve this.

However, when looking at the interaction with obfuscated instead of attitude-confirming search results, we found that both targeted and random obfuscations of search results were effective in decreasing the proportion of clicks on these obfuscated search results. This implies that search result obfuscations are a powerful tool in steering users' search result selection behavior, which consequently could be misused for purposes other than the users' benefit, raising ethical considerations which we follow up on in Section 5.1. This observation could be explained in two ways: either (1) participants blindly trusted the system's decision, hindering them from realizing that in the random obfuscation condition the results obfuscated were indeed random (and not attitude-confirming), or (2) they simply ignored the obfuscated search results and focused on clearly visible search results that did not requiring additional effort. In the targeted obfuscation condition these were to a high proportion (75%) attitude-opposing. The second explanation is in line with the findings of Kaiser et al. [22] that warnings are more effective in decreasing interaction with an item when they require user interaction, partly due to the additional effort introduced to the users' workflow. They warned that this might decrease user experience, not foster informed decision making, and result in habituation effects.

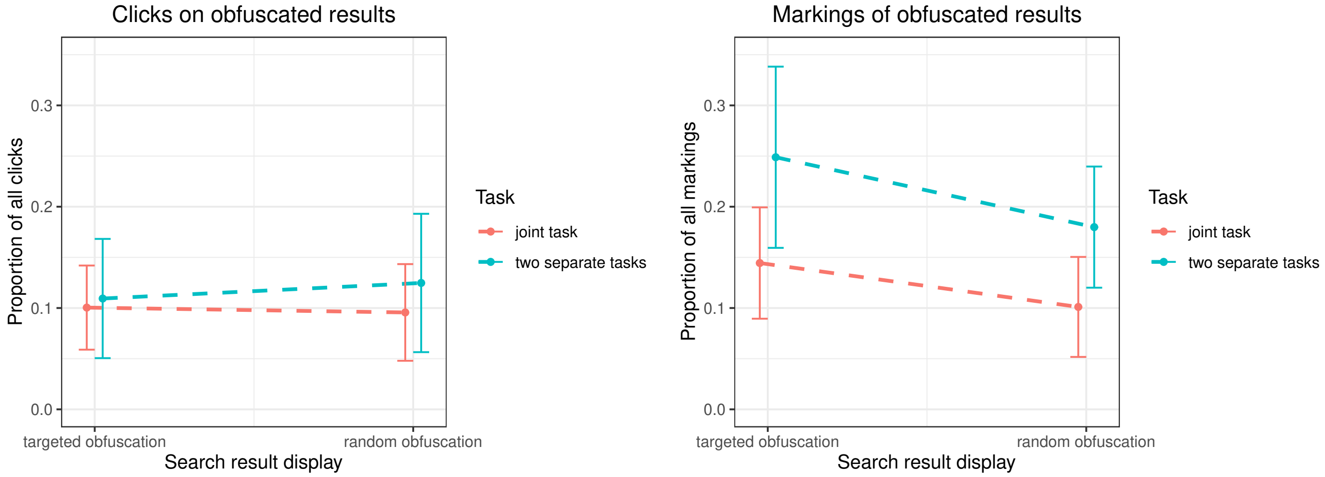


Figure 4: Interaction with obfuscated search results: mean proportion of participant’s clicks on (left) and markings of (right) obfuscated search results in the targeted obfuscation and random obfuscation condition and per task condition (joint task and two separate tasks) with 95% confidence intervals. A proportion of 0.1 implies that 10% of a participant’s clicks/markings were on obfuscated search results.

Further, we consider our findings in light of the *Elaboration Likelihood Model* (ELM) [43], which distinguishes between a *central* and a *peripheral* route to persuasion. It seems likely that the effect we observed was caused by the *peripheral* route (i.e., interacting with attitude-opposing search results because interaction requires less effort and is thus more attractive). The authors stated that attitude change caused by peripheral instead of central cues is less enduring, relatively temporary, and unpredictive of behavior. An approach applying central cues was investigated by Hube et al. [20]. They found that in a setting with no option of choosing a path of lower effort, warnings which require user interaction were effective in mitigating worker bias and improving performance. Thus, for a setting in which users can choose a path of lower effort, we should strive to find an effective combination of peripheral cues, to guide users’ interaction and to catch their attention, and central cues, to motivate careful and analytic consideration of information.

Effect of task. We did not find any evidence for an effect of task on the proportion of attitude-confirming markings. This implies that repeated exposure to the warning label (two separate tasks), does not alter the effect of the obfuscation with warning label on markings of attitude-confirming search results, at least to the limited extent to which we were able to observe this in a single session experiment. However, we observed that participants in the joint task condition tended to spend a longer time on the SERP and to click on more search results than participants in the two separate tasks condition. While the former observation might be explained by participants in the joint task condition doing two tasks (clicking and marking) instead of one (only clicking), the latter suggests that the marking task might have motivated increased clicking on search results. Further research on the potential effect of task design on confirmation bias and analytical information processing is required.

Exploratory: Interaction of task and display. During exploratory analysis, we furthermore observed a higher proportion of markings of *obfuscated* search results in the two separate task condition than in the joint task condition for targeted and random obfuscation. This observation might be explained by a task design decision for the separate task condition (see Section 5.3): recall that in the two separate tasks condition, during the marking task, search results were not obfuscated but merely displayed with the warning label (see Section 3.1.3). In the joint task condition, however, clicking and marking was done in a single task and obfuscated search results remained obfuscated, unless participants actively clicked the button to view the search result. Removing obfuscation (two separate tasks) seemed to result in more interaction, suggesting that obfuscation is effective in decreasing interaction with search results. This observation further supports the explanation that participants might have chosen the path of lowest effort [22] due to peripheral cues of persuasion for behavioral change [43] instead of carefully and analytically considering the information. Thus, follow-up research could be tailored specifically to answer the question of what causes decreased interaction with search results that are obfuscated with a warning label.

Effect of user-related behavioral patterns. We found evidence for a correlation of participants’ proportion of attitude-confirming clicks and markings. This finding suggests that behavioral patterns, caused by user-related factors which influence the interaction with search results of different viewpoints, exist. Hence, further research is required to investigate how situational or stable factors which have been found to moderate users’ search behavior, such as attitude strength [27], interest in the topic [50], and personality traits (e.g. Need for Cognition [54]), might affect the effectiveness of confirmation bias mitigation approaches and how they should be adapted accordingly.

Considerations for real-world applications. Collecting viewpoint annotations for a handpicked selection of 200 search results and specifically asking participants for their attitude on the four selected topics was necessary for conducting our controlled user study but limits the applicability of our approach to complex real-world scenarios. Enabling effective real-world applications of confirmation bias mitigation strategies in search may however be possible by drawing from related research. For instance, recent advances on automatic stance detection [1, 29] and perspective discovery [10] provide important tools towards assigning correct viewpoint labels automatically. Furthermore, approaches for automatically measuring viewpoint diversity in search results [11] or real-time confirmation bias detection, as researched for example for the field of visual analytics [56], might prove useful here. Real-world application of the confirmation bias mitigation approach investigated with this work will lend itself more to large-scale implementation as such tools become more advanced. However, our findings urge us to exercise caution when going about a real-world implementation of such approaches due to a number of ethical considerations which we discuss in the following section.

5.1 Ethical Considerations

The two potential explanations for no differences in the interaction behavior with obfuscated search results between targeted and random obfuscation condition raise ethical concerns with regard to using obfuscations with warnings for confirmation bias mitigation during web search: (1) If the findings can be explained by high trust in warning labels (i.e., even if they are applied incorrectly as was the case in the random obfuscation condition), this would allow for exploitation and misuse for someone's interests and against the user's benefit. (2) If, on the other hand, the findings can be explained by the users' (potentially unintentional) ignorance of or blindness for obfuscated search results and their tendency not to engage with search results if engagement requires additional effort (i.e., clicking one button), then we would be battling cognitive bias by harnessing other cognitive biases. This is effective in getting users to interact with attitude-opposing search results but most likely is not an appropriate approach to motivate analytic information processing. Consequently, this approach would threaten user autonomy and thus not fulfill the requirements for *nudges to reason* stated by [30]. An improvement could be to design obfuscations with warning labels more saliently so that it is less likely that users unintentionally ignore them. However, this proposal requires further research.

Additional ethical considerations concern the practical implementation of approaches of confirmation bias mitigation during search. As discussed in the previous section, users' attitude on a topic would have to be elicited automatically from their search behavior. However, automatically eliciting personal attitudes on different topics, including sensitive personal information such as political beliefs, requires user-data collection and storing that is not compatible with GDPR regulations. Thus we promote approaches which base the decision on what to obfuscate merely on the observed behavior in a single search session. This could be done for example by applying targeted obfuscations after a user has selected a number of articles all supporting the same viewpoint. Approaches of real-time confirmation bias detection during search which do not require storing sensitive user-data need to be examined further in future studies.

Based on these considerations, we propose the following **ethical guidelines**:

(1) Apply obfuscations for confirmation bias mitigation exclusively to the users' benefit.
(2) Obtain users' consent before obfuscating to mitigate confirmation bias and enable consent withdrawal.
(3) Explain transparently why an item is obfuscated so that users understand the system's decision and are able to detect system errors (i.e. incorrect obfuscations).
(4) Include simple mechanisms that allow users to control/correct the obfuscation feature if necessary due to incorrect system decisions or as desired by the user.

5.2 Implications and Design guidelines

From our findings we learn that obfuscations with warning labels, requiring additional effort to view a search result, are an effective approach to decrease interaction with search results, whereas our exploratory findings suggest that the search result obfuscation might have had a greater impact than the warning label. If such obfuscations are applied targeting attitude-confirming search results, they can effectively nudge users to interact with a higher proportion of attitude-opposing search results than they would do without obfuscations. Thus, if applied carefully, this approach might help users to overcome confirmation bias while selecting search results during online search.

Our findings have practical implications for the implementation of obfuscation-based approaches for confirmation bias mitigation during search, thus we formulated the following **design guidelines**:

(1) Design obfuscation in a way that requires an appropriate amount of additional effort to view the item (i.e. button to actively accept the risk of confirmation bias). *Given our findings it seems that users are likely to take the path of lowest effort and thus interact less with items that would require additional effort. However, according to Kaiser et al. [22] the amount needs to be selected diligently to avoid decreasing user experience and warning fatigue.*
(2) Select diligently what to obfuscate (target attitude-confirming search results). *Our findings show that users interact less with obfuscated search results no matter if obfuscation is targeted or random. Thus which items to obfuscate should be decided carefully and, if necessary, be adapted.*
(3) Design obfuscations with warning labels with an appropriate level of salience to avoid unintentional ignorance of or blindness for the obfuscated items which would threaten user autonomy. *Our research design failed to detect if the cause of less interaction with obfuscated search results might have been users' ignorance of, or blindness for these search results. However, to fulfill the requirements for ethical permissible nudges that do not threaten user autonomy stated by [30], it is important to ensure that obfuscated search results are not unintentionally ignored because they did not capture users' attention.*

5.3 Limitations and future work

To be able to conduct a controlled user study, we had to construct an artificial scenario, for which we pre-selected the topics and search results. Even though we assigned each participant to one of the topics for which they reported to have the strongest attitude, they still might not have had great interest in the topic, or, as formulated by [5], they had "no skin in the game". Further, the setup did not allow participants to formulate one or multiple own queries and to conduct the search, as they would naturally do, but they were forced to interact exclusively with the 12 pre-selected search results. However, we attempted to make the task as realistic and relatable as possible and refrained from enforcing minimum time requirements, even though this meant excluding data.

Another limitation of this study is that we only observed one single search session, exposing participants to obfuscations with warning labels for a limited time. Yet, most of us use search engines multiple times per day and thus are exposed to the search engine interface very frequently and adapt our behavior according to our intentions and the search engine's features. It would therefore be interesting to observe user behavior and potential adaptions to warning labels and obfuscations in a less controlled and more natural setting, and over a longer period of time.

Lastly, we did not investigate the effects of warning label and obfuscation independently and systemically. We did, however, display search results merely with a warning label (not obfuscated) during the marking task for the two separate tasks condition and compared the interaction behavior to the joint task condition, in which search results were obfuscated with a warning label. Yet, this decision constitutes another limitation of this study, since it was based on attempting to control for and investigate effects of multiple exposure to both the warning labels and the search results, as discussed previously. Ultimately, this design decision prevents us from drawing valid conclusions on the effect of multiple exposure to the warning labels, but unveils that obfuscations with warning labels might have been more effective than merely warning labels in decreasing interaction with search results. Targeted investigation of the effects of warning label and obfuscation independently needs to be done in future studies.

6 CONCLUSION

We presented a user study investigating the effect of obfuscations with warning labels about confirmation bias, on the interaction with viewpoint-annotated search results on debated topics. We found that obfuscations result in decreased interaction with search results and that targeted obfuscations of attitude-confirming search results are effective in increasing the interaction with attitude-opposing search results. However, it remains to be clarified whether this effect was observed because participants trusted the warning label or avoided additional effort and ignored the obfuscated search results. Given these findings, we call for strict regulations, allowing an application of search result obfuscations exclusively to the users' benefit, with their consent, and in a transparent and controllable way.

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

[1] Abeer ALDayel and Walid Magdy. 2021. Stance Detection on Social Media: State of the Art and Trends. *Information Processing & Management* 58 (July 2021), 102597. https://doi.org/10.1016/j.ipm.2021.102597

[2] Jisun An, Daniele Quercia, Meeyoung Cha, Krishna Gummadi, and Jon Crowcroft. 2014. Sharing Political News: The Balancing Act of Intimacy and Socialization in Selective Exposure. *EPJ Data Science* 3 (Dec. 2014), 12. https://doi.org/10.1140/epjds/s13688-014-0012-2

[3] Evangelia Anagnostopoulou, Babis Magoutas, Efthimios Bothos, Johann Schrammel, Rita Orji, and Gregoris Mentzas. 2017. Exploring the Links Between Persuasion, Personality and Mobility Types in Personalized Mobility Applications. In *Persuasive Technology: Development and Implementation of Personalized Technologies to Change Attitudes and Behaviors*, Peter W. de Vries, Harri Oinas-Kukkonen, Liseth Siemons, Nienke Beerlage-de Jong, and Lisette van Gemert-Pijnen (Eds.). Vol. 10171. Springer International Publishing, Cham, 107–118. https://doi.org/10.1007/978-3-319-55134-0_9

[4] Jennifer J. Argo and Kelley J. Main. 2004. Meta-Analyses of the Effectiveness of Warning Labels. *Journal of Public Policy & Marketing* 23 (Sept. 2004), 193–208. https://doi.org/10.1509/jppm.23.2.193.51400

[5] Leif Azzopardi. 2021. Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. ACM, Canberra ACT Australia, 27–37. https://doi.org/10.1145/3406522.3446023

[6] John T. Cacioppo, Richard E. Petty, and Katherine J. Morris. 1983. Effects of Need for Cognition on Message Evaluation, Recall, and Persuasion. *Journal of Personality and Social Psychology* 45 (1983), 805–818. https://doi.org/10.1037/0022-3514.45.4.805

[7] Noel Carroll. 2014. In Search We Trust. *International Journal of Knowledge Society Research* 5, 1 (2014), 12–27. https://doi.org/10.4018/ijksr.2014010102

[8] Katherine Clayton, Spencer Blair, Jonathan A. Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, Morgan Sandhu, Rachel Sang, Rachel Scholz-Bright, Austin T. Welch, Andrew G. Wolff, Amanda Zhou, and Brendan Nyhan. 2020. Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media. *Political Behavior* 42 (Dec. 2020), 1073–1095. https://doi.org/10.1007/s11109-019-09533-0

[9] Arthur R. Cohen, Ezra Stotland, and Donald M. Wolfe. 1955. An Experimental Investigation of Need for Cognition. *The Journal of Abnormal and Social Psychology* 51 (1955), 291–294. https://doi.org/10.1037/h0042761

[10] Tim Draws, Jody Liu, and Nava Tintarev. 2020. Helping users discover perspectives: Enhancing opinion mining with joint topic models. In *2020 International Conference on Data Mining Workshops (ICDMW)*. IEEE, Sorrento, Italy, 23–30. https://doi.org/10.1109/ICDMW51313.2020.00013

[11] Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. 2021. Assessing Viewpoint Diversity in Search Results Using Ranking Fairness Metrics. *ACM SIGKDD Explorations Newsletter* 23, 1 (May 2021), 50–58. https://doi.org/10.1145/3468507.3468515

[12] Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. 2021. This Is Not What We Ordered: Exploring Why Biased Search Result Rankings Affect User Attitudes on Debated Topics. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Virtual Event Canada, 295–305. https://doi.org/10.1145/3404835.3462851

[13] John Fox and Sanford Weisberg. 2019. *An R Companion to Applied Regression* (third ed.). Sage, Thousand Oaks CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

[14] Gerd Gigerenzer. 2008. Why Heuristics Work. *Perspectives on Psychological Science* 3 (Jan. 2008), 20–29. https://doi.org/10.1111/j.1745-6916.2008.00058.x

[15] 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*. 228–240.

[16] Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures* 1 (April 2007), 77–89. https://doi.org/10.1080/19312450709336664

[17] Thomas T Hills. 2019. The Dark Side of Information Proliferation. *Perspectives on Psychological Science* 14 (2019), 323–330.

[18] Adrian Holzer, Nava Tintarev, Samuel Bendahan, Bruno Kocher, Shane Greenup, and Denis Gillet. 2018. Digitally Scaffolding Debate in the Classroom. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–6. https://doi.org/10.1145/3170427.3188499

[19] Steven Houben and Christian Weichel. 2013. Overcoming Interaction Blindness through Curiosity Objects. In *CHI '13 Extended Abstracts on Human Factors in*

Computing Systems on - CHI EA ’13. ACM Press, Paris, France, 1539. https://doi.org/10.1145/2468356.2468631

[20] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–12. https://doi.org/10.1145/3290605.3300637

[21] Mathias Jesse and Dietmar Jannach. 2021. Digital Nudging with Recommender Systems: Survey and Future Directions. *Computers in Human Behavior Reports* 3 (Jan. 2021), 100052. https://doi.org/10.1016/j.chbr.2020.100052

[22] Ben Kaiser, Jerry Wei, Elena Lucherini, Kevin Lee, J. Nathan Matias, and Jonathan Mayer. 2021. Adapting Security Warnings to Counter Online Disinformation. In *30th {USENIX} Security Symposium ({USENIX} Security 21)*.

[23] Alboukadel Kassambara. 2021. *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*. https://CRAN.R-project.org/package=rstatix R package version 0.7.0.

[24] Varol Kayhan. 2015. Confirmation Bias: Roles of Search Engines and Search Contexts. (2015), 18.

[25] Seongho Kim. 2015. *ppcor: Partial and Semi-Partial (Part) Correlation*. https://CRAN.R-project.org/package=ppcor R package version 1.1.

[26] Jan Kirchner and Christian Reuter. 2020. Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness. *Proceedings of the ACM on Human-Computer Interaction* 4 (Oct. 2020), 1–27. https://doi.org/10.1145/3415211

[27] Silvia Knobloch-Westerwick and Jingbo Meng. 2009. Looking the Other Way: Selective Exposure to Attitude-Consistent and Counterattitudinal Political Information. *Communication Research* 36 (June 2009), 426–448. https://doi.org/10.1177/0093650209333030

[28] Silvia Knobloch-Westerwick, Benjamin K. Johnson, and Axel Westerwick. 2015. Confirmation Bias in Online Searches: Impacts of Selective Exposure Before an Election on Political Attitude Strength and Shifts. *Journal of Computer-Mediated Communication* 20 (2015), 171–187. https://doi.org/10.1111/jcc4.12105

[29] Dilek Küçük and Fazli Can. 2020. Stance Detection: A Survey. *Comput. Surveys* 53 (May 2020), 1–37. https://doi.org/10.1145/3369026

[30] Neil Levy. 2017. Nudges in a Post-Truth World. *Journal of medical ethics* 43 (2017), 495–500.

[31] Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest* 13 (Dec. 2012), 106–131. https://doi.org/10.1177/1529100612451018

[32] 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*. 184–196.

[33] Scott O Lilienfeld, Rachel Ammirati, and Kristin Landfield. 2009. Giving Debiasing Away: Can Psychological Research on Correcting Cognitive Errors Promote Human Welfare? *Perspectives on psychological science* 4 (2009), 390–398.

[34] Paul Mena. 2020. Cleaning Up Social Media: The Effect of Warning Labels on Likelihood of Sharing False News on Facebook. *Policy & Internet* 12 (2020), 165–183. https://doi.org/10.1002/poi3.214

[35] Microsoft. 2021. Web Search API: Microsoft Bing. https://www.microsoft.com/en-us/bing/apis/bing-web-search-api

[36] Martijn Millecamp, Robin Haveneers, and Katrien Verbert. 2020. Cogito Ergo Quid? The Effect of Cognitive Style in a Transparent Mobile Music Recommender System. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP ’20)*. Association for Computing Machinery, New York, NY, USA, 323–327. https://doi.org/10.1145/3340631.3394871

[37] Fauzan Misra. 2019. Accountability pressure as debiaser for confirmation bias in information search and tax consultant’s recommendations. *Journal of Indonesian Economy and Business* 34 (July 2019), 80. https://doi.org/10.22146/jieb.40019

[38] Sean A Munson and Paul Resnick. 2010. Presenting Diverse Political Opinions: How and How Much. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1457–1466.

[39] Philip M. Napoli. 1999. Deconstructing the diversity principle. *Journal of Communication* 49, 4 (1999), 7–34. https://doi.org/10.1111/j.1460-2466.1999.tb02815.x

[40] Raymond S Nickerson. 1998. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. (1998), 46.

[41] Derek H. Ogle, Powell Wheeler, and Alexis Dinno. 2021. *FSA: Fisheries Stock Analysis*. https://github.com/droglenc/FSA R package version 0.8.32.

[42] Gordon Pennycook and David G. Rand. 2019. Lazy, Not Biased: Susceptibility to Partisan Fake News Is Better Explained by Lack of Reasoning than by Motivated Reasoning. *Cognition* 188 (July 2019), 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

[43] Richard E Petty and John T Cacioppo. 19986. The Elaboration Likelihood Model of Persuasion. (19986), 1–24.

[44] ProCon.org. 2021. Homepage. https://www.procon.org/

[45] Prolific. 2021. Quickly find research participants you can trust. https://www.prolific.co/

[46] Qualtrics. 2021. Qualtrics XM - Experience Management Software. https://www.qualtrics.com/

[47] R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[48] Arno J. Rethans, John L. Swasy, and Lawrence J. Marks. 1986. Effects of Television Commercial Repetition, Receiver Knowledge, and Commercial Length: A Test of the Two-Factor Model. *Journal of Marketing Research* 23 (Feb. 1986), 50–61. https://doi.org/10.1177/002224378602300106

[49] Alisa Rieger, Mariët Theune, and Nava Tintarev. 2020. Toward Natural Language Mitigation Strategies for Cognitive Biases in Recommender Systems. In *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*. Association for Computational Linguistics, Dublin, Ireland, 50–54.

[50] Josephine B. Schmitt, Christina A. Debbelt, and Frank M. Schneider. 2018. Too Much Information? Predictors of Information Overload in the Context of Online News Exposure. *Information, Communication & Society* 21 (Aug. 2018), 1151–1167. https://doi.org/10.1080/1369118X.2017.1305427

[51] Jack B Soll, Katherine L Milkman, and John W Payne. 2015. A User’s Guide to Debiasing. *The Wiley Blackwell handbook of judgment and decision making* 2 (2015), 924–951.

[52] Richard H. Thaler, Cass R. Sunstein, and John P. Balz. 2010. *Choice Architecture*. SSRN Scholarly Paper. Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.1583509

[53] Rob Tieben, Tilde Bekker, and Ben Schouten. 2011. Curiosity and Interaction: Making People Curious through Interactive Systems. In *Proceedings of HCI 2011 The 25th BCS Conference on Human Computer Interaction*. https://doi.org/10.14236/ewic/HCI2011.66

[54] Yariv Tsfati and Joseph N. Cappella. 2005. Why Do People Watch News They Do Not Trust? The Need for Cognition as a Moderator in the Association Between News Media Skepticism and Exposure. *Media Psychology* 7 (Aug. 2005), 251–271. https://doi.org/10.1207/S1532785XMEP0703_2

[55] Amazon Mechanical Turk. 2021. https://www.mturk.com/

[56] Emily Wall, Leslie M Blaha, Lyndsey Franklin, and Alex Endert. 2017. Warning, Bias May Occur: A Proposed Approach to Detecting Cognitive Bias in Interactive Visual Analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 104–115.

[57] Hadley Wickham. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

[58] Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. 2020. *dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr R package version 1.0.2.

[59] Wikipedia. 2021. Confirmation bias. https://en.wikipedia.org/wiki/Confirmation_bias

Visual-Meta Appendix

The data below is what we call Visual-Meta. It is an approach to add information about a document to the document itself, on the same level of the content (in style of BibTeX).
It is very important to make clear that Visual-Meta is an approach more than a specific format and that it is based on wrappers. Anyone can make a custom wrapper for custom metadata and append it by specifying what it contains: for example @dublin-core or @rdfs.
The way we have encoded this data, and which we recommend you do for your own documents, is as follows:
When listing the names of the authors, they should be in the format 'last name', a comma, followed by 'first name' then 'middle name' whilst delimiting discrete authors with ('and') between author names, like this: Shakespeare, William and Engelbart, Douglas C.
Dates should be ISO 8601 compliant.
Every citable document will have an ID which we call 'vm-id'. It starts with the date and time the document's metadata/Visual-Meta was 'created' (in UTC), then max first 10 characters of document title.
To parse the Visual-Meta, reader software looks for Visual-Meta in the PDF by scanning the document from the end, for the tag @{visual-meta-end}. If this is found, the software then looks for @{visual-meta-start} and uses the data found between these tags. This was written September 2021. More information is available from https://visual-meta.info for as long as we can maintain the domain.

@{visual-meta-start}

@{visual-meta-header-start}
@visual-meta{version = {1.1},
generator = {ACM Hypertext 21},
organisation = {Association for Computing Machinery}, }

@{visual-meta-header-end}

@{visual-meta-bibtex-self-citation-start}

@inproceedings{10.1145/3465336.3475101,
author = {Rieger, Alisa and Draws, Tim and Theune, Mariët and Tintarev, Nava},
title = {This Item Might Reinforce Your Opinion: Obfuscation and Labeling of Search Results to Mitigate Confirmation Bias},
year = {2021},
isbn = {978-1-4503-8551-0},
publisher = {Association for Computing Machinery},
address = {New York, NY, USA},
url = {https://doi.org/10.1145/3465336.3475101},
doi = {10.1145/3465336.3475101},
abstract = {During online information search, users tend to select search results that confirm previous beliefs and ignore competing possibilities. This systematic pattern in human behavior is known as confirmation bias. In this paper, we study the effect of obfuscation (i.e., hiding the result unless the user clicks on it) with warning labels and the effect of task on interaction with attitude-confirming search results. We conducted a preregistered, between-subjects crowdsourced user study (N=328) comparing six groups: three levels of obfuscation (targeted, random, none) and two levels of task (joint, two separate) for four debated topics. We found that both types of obfuscation influence user interactions, and in particular that targeted obfuscation helps decrease interaction with attitude-confirming search results. Future work is needed to understand how much of the observed effect is due to the strong influence of obfuscation, versus the warning label or the task design. We discuss design guidelines concerning system goals such as decreasing consumption of attitude-confirming search results, versus nudging users toward a more analytical mode of information processing. We also discuss implications for future work, such as the effects of interventions for confirmation bias mitigation over repeated exposure. We conclude with a strong word of caution: measures such as obfuscations should only be used for the benefit of the user, e.g., when they explicitly consent to mitigating their own biases.},
numpages = {11},
keywords = {Confirmation Bias; Web Search; Warning Labels; Obfuscation; Nudging; Cognitive Bias Mitigation},
location = {Virtual Event, USA},
series = {HT '21},
vm-id = {10.1145/3465336.3475101} }

@{visual-meta-bibtex-self-citation-end}

@{visual-meta-end}