Nudges to Mitigate Confirmation Bias during Web Search on Debated Topics Support vs. Manipulation

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

DOI
10.1145/3635034

Publication date
2024

Document Version
Final published version

Published in
ACM Transactions on the Web

Nudges to Mitigate Confirmation Bias during Web Search on Debated Topics: Support vs. Manipulation

ALISA RIEGER and TIM DRAWS, Delft University of Technology, Netherlands
MARIËT THEUNE, University of Twente, Netherlands
NAVA TINTAREV, Maastricht University, Netherlands

When people use web search engines to find information on debated topics, the search results they encounter can influence opinion formation and practical decision-making with potentially far-reaching consequences for the individual and society. However, current web search engines lack support for information-seeking strategies that enable responsible opinion formation, e.g., by mitigating confirmation bias and motivating engagement with diverse viewpoints. We conducted two preregistered user studies to test the benefits and risks of an intervention aimed at confirmation bias mitigation. In the first study, we tested the effect of warning labels, warning of the risk of confirmation bias, combined with obfuscations, hiding selected search results per default. We observed that obfuscations with warning labels effectively reduce engagement with search results. These initial findings did not allow conclusions about the extent to which the reduced engagement was caused by the warning label (reflective nudging element) versus the obfuscation (automatic nudging element). If obfuscation was the primary cause, this would raise concerns about harming user autonomy. We thus conducted a follow-up study to test the effect of warning labels and obfuscations separately.

According to our findings, obfuscations run the risk of manipulating behavior instead of guiding it, while warning labels without obfuscations (purely reflective) do not exhaust processing capacities but encourage users to actively choose to decrease engagement with attitude-confirming search results. Therefore, given the risks and unclear benefits of obfuscations and potentially other automatic nudging elements to guide engagement with information, we call for prioritizing interventions that aim to enhance human cognitive skills and agency instead.

CCS Concepts: • **Human-centered computing** → **User studies**; **User centered design; • Information systems** → **Search interfaces**;

Additional Key Words and Phrases: Web search, debated topics, nudging, cognitive bias mitigation, cognitive reflection

ACM Reference format:
Alisa Rieger, Tim Draws, Mariët Theune, and Nava Tintarev. 2024. Nudges to Mitigate Confirmation Bias during Web Search on Debated Topics: Support vs. Manipulation. *ACM Trans. Web* 18, 2, Article 27 (March 2024), 27 pages.
https://doi.org/10.1145/3635034

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.

Authors' addresses: A. Rieger and T. Draws, Delft University of Technology, Van Mourik Broekmanweg 6, 2628 CD Delft, Zuid-Holland, Netherlands; e-mails: {a.rieger, t.a.draws}@tudelft.nl; M. Theune, Hallenweg 15, 7522 NH, Enschede, Overijssel, Netherlands; e-mail: m.theune@utwente.nl; N. Tintarev, Maastricht Univeristy, Paul-Henri Spaaklaan 1, 6229 EN Maastricht, Limburg, Netherlands; e-mail: n.tintarev@maastrichtuniversity.nl.

1559-1131/2024/03-ART27
https://doi.org/10.1145/3635034

1 INTRODUCTION

Web search engines have evolved into tools that are used to satisfy all kinds of information needs, many of them more complex than simple *lookup* tasks that have been the main focus of information retrieval research [22, 38, 57, 68]. For instance, people use web search engines to find information before forming opinions that can lead to practical decisions. Such decisions can have different levels of impact, ranging from trivial day-to-day (e.g., what movie to watch) to important and consequential (e.g., whether to get a vaccine) [9]. Searches that lead to decisions with high impact on the individual decision maker and/or society often concern *debated topics*, subjects of ongoing discussion such as *whether to become vegan* [17] or *whom to vote for* [14]. Forming such opinions responsibly would require thorough and unbiased information seeking [31, 45].

The required information-seeking strategies are known to be cognitively demanding [70] and are not sufficiently supported by current web search engines [38, 57, 60]. To alleviate cognitive demand, searchers might tend to adopt biased search behaviors, such as favoring information that aligns with prior beliefs and values while disregarding conflicting information *(confirmation bias)* [2, 44]. For promoting thorough and unbiased search behavior, search engines could employ behavioral interventions [32, 37]. In the context of search on debated topics, such interventions could aim at mitigating confirmation bias, for instance by decreasing engagement with attitude-confirming search results. Interventions to decrease engagement with selected items have attracted substantial research attention within an alternate context, namely of countering the spread of misinformation [33]. An approach that has shown notable success consists of *warning labels* and *obfuscations* to flag and decrease the ease of access to items that likely contain misinformation [10, 29, 40]. This intervention combines both transparent reflective and transparent automatic nudging elements [23] (see Figure 1): It *prompts reflective choice* by presenting warning labels and it *influences behavior* by decreasing the ease of access to the item through default obfuscations [8].

The parallels between misinformation and confirmation bias regarding objectives of interventions (i.e., decreased engagement with targeted items), and underlying cognitive processes that increase the susceptibility (i.e., lack of analytical thinking) [2, 46], motivated us to investigate with the **warning label and obfuscation study** (see Section 3) whether this intervention is likewise successful for confirmation bias mitigation and in supporting thorough information-seeking during search on debated topics. By applying warning labels and obfuscations (see Figure 2) to attitude-confirming search results (i.e., search results that express a viewpoint in line with the user's pre-search opinion on the topic), we aimed at encouraging users with strong prior attitudes to engage with different viewpoints and gain a well-rounded understanding of the topic. To understand the benefits and risks of this intervention, we conducted two user studies (see Figure 3). With the first, we investigated the effect of warning labels and obfuscations combined (**warning label and obfuscation study**) on searchers' confirmation bias. Subsequently, we conducted a follow-up study in which we investigated how different searchers and their search behavior are affected by warning labels and obfuscations separately (**automatic vs. reflective study**).

In the *warning label and obfuscation study* detailed in Section 3, we investigated the effect of the intervention on searchers' confirmation bias. The results show that warning labels and obfuscations effectively decrease interaction with targeted search results ($f = 0.35$). Yet, this first study did not provide sufficient grounds for determining what specifically caused the observed effect; i.e., whether **(1)** participants read the warning label (reflective nudging element) and, now aware of confirmation bias, actively decided to interact less with attitude confirming search results, or **(2)** participants took the path of lowest effort and unconsciously ignored all obfuscated items (automatic nudging element), since interaction with those required increased effort. Exploratory insights from this study indicate that the extent to which the reflective or automatic elements of the intervention caused the effect might vary across users with distinct cognitive styles. *Cognitive*

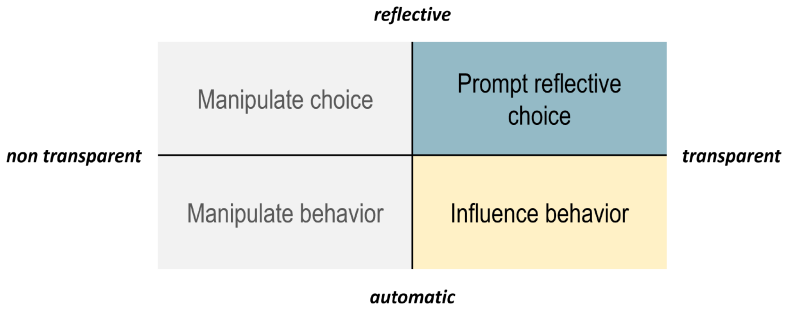

Fig. 1. Categories of nudging elements, adapted from Hansen and Jespersen [23]. Since this work investigates interventions that aim at guiding (as opposed to manipulating) user behavior, we only consider nudges from the transparent categories.

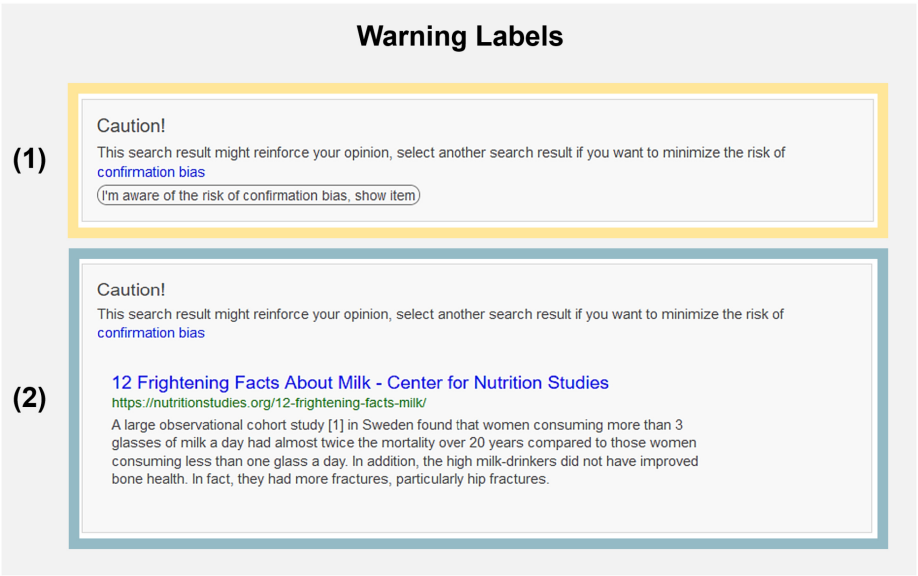

Fig. 2. Warning labels. (1) Warning label with obfuscation, after participants clicked on *show-button*, the search result was revealed and they saw (2) Warning label without obfuscation. In the *warning label without obfuscation* conditions in the automatic vs. reflective study, the default shown to participants was condition (2).

style describes an individual's tendency to rely more on analytic, effortful or intuitive, effortless thinking [6, 15]. Furthermore, the exploratory observations suggest that both the search result display and the individuals' cognitive style might impact the searcher beyond their search interactions, namely their attitude change and awareness of bias.

In an effort to better understand what caused the effect of decreased interaction and how the interventions impact searchers with distinct cognitive style, we initiated the **automatic vs. reflective study** as a follow-up. With the **automatic vs. reflective study** detailed in Section 4, we tested the effect of the reflective element of the intervention separately by adding a search result display condition (for an overview of the conditions, see Figure 3) in which search results were displayed with the warning label (reflective) but not obfuscated (see (2) in Figure 2).

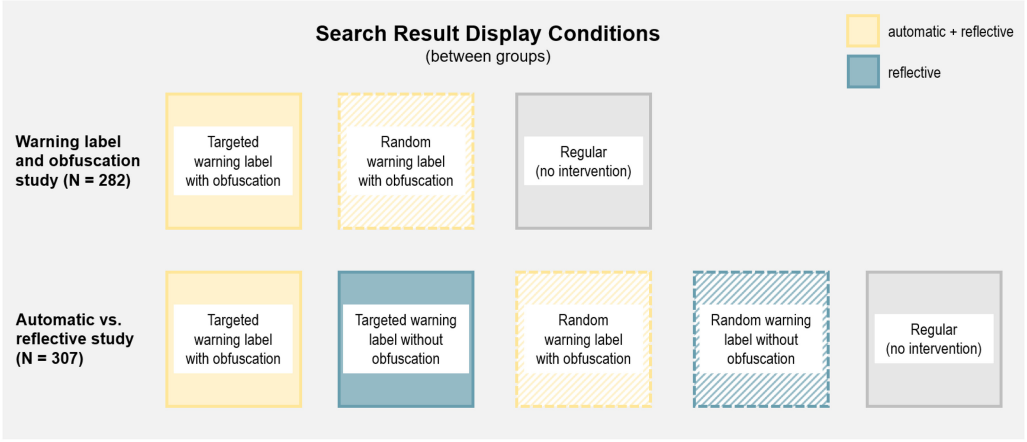

Fig. 3. Search result display conditions in the warning label and obfuscation study (top) and automatic vs. reflective study (bottom).

The **automatic vs. reflective study** replicated the finding of a moderate effect of obfuscations with warning labels that reduced clicks on attitude-confirming search results for a new set of search results ($f = 0.30$). Moreover, we observed that warning labels without obfuscation (reflective) reduce engagement when applied to attitude-confirming search results, but, in contrast to warning labels with obfuscation, do not reduce engagement when applied to randomly selected search results. Thus, our key takeaways from both studies are that obfuscations, and possibly other automatic nudging elements, run the risk of manipulating behavior instead of guiding it while warning labels without obfuscations effectively encourage users to choose to engage with less attitude-confirming search results.

With this article, we make the following contributions:

—Discussion of benefits and risks of warning labels and obfuscations to mitigate confirmation bias among diverse users informed by exploratory insights from a preregistered user study with 282 participants (main findings published in [54]) and findings of a preregistered follow-up study with 307 participants;
—Design implications for behavioral interventions that aim at supporting responsible opinion formation during web search;
—Validation of findings on the effect of warning labels with obfuscations on confirmation bias (published in [54]) through replication;
—Two datasets with interaction data and questionnaire responses, publicly available in the repositories at links in Footnotes 1 and 9.

2 RELATED WORK

In this section, we discuss background literature on different related areas of research. These include search on debated topics and confirmation bias, interventions to guide web interactions, and the role of cognitive reflection during engagement with information.

2.1 Search on Debated Topics and Confirmation Bias

Individuals may turn to web search to develop or revise their opinions on different subject matters, e.g., to satisfy individual interest or to gather advice before making decisions [9, 68]. This

can concern *debated topics*, subjects on which individuals or groups have different opinions, for instance, due to conflicting values, competing interests, and various possible perspectives from which to view the issues. Web search on debated topics can be consequential for both individuals and society at large, given its potential to influence practical decision-making [9, 14, 39]. Thus, we are interested in how web search could support people in forming opinions responsibly.

The notion of *responsibility* in opinion formation has been thoroughly discussed by philosophers in the field of epistemology [31, 45]. Kornblith [31], for instance, reasons that responsible beliefs are the product of actively gathering evidence and critically evaluating it. For responsible opinion formation, individuals should thus gather information to gain a well-rounded understanding of the topic and the various arguments and form opinions and make decisions based on the synthesized information they gathered and knew before. Traditionally, the objective of gaining a well-rounded understanding of the topic and arguments could be supported by (public) media and news outlets which are subject to regulations and ethical guidelines, e.g., regarding quality and diversity of content [24]. However, rather than primarily consulting curated journalistic content, people increasingly rely on search engines to actively search for information on debated topics to form opinions or make decisions [9, 68]. The opaque nature of search engines that automatically filter and rank resources and are not (yet) bound to follow principles of responsible information proliferation (e.g., exposure diversity [24]) can prevent users from recognizing whether the provided information is complete and reliable [41, 60]. Web search for responsible opinion formation thus requires self-reliant, thorough, exploratory search behavior, which is known to be cognitively demanding [26, 48, 52].

As a means of simplifying complex search tasks, searchers are prone to resort to heuristics and systematic shortcuts [2]. While such shortcuts typically lead to more efficient actions and decisions under constraint resources (e.g., information-processing capacities or time) [19], they can result in cognitive biases – systematic errors in judgment and decision-making [65]. A prevailing strategy to limit the cognitive demand of search tasks is the *confirmation bias*, the human tendency to prioritize information that confirms prior attitudes [44]. Confirmation bias thus impedes engagement with diverse viewpoints and can manifest throughout the various stages of the information search procedure: It can cause users to employ affirmative testing techniques while querying, interact mainly with search results that align with their attitudes, and disregard information that counters their attitude when evaluating arguments to form beliefs or make decisions [2, 66, 69, 73]. Yet, search engines could be designed to accommodate more complex and exploratory search tasks and support thorough and unbiased information-seeking strategies [57, 60].

2.2 Guiding Web Interactions

To empower individuals online, Lorenz-Spreen et al. [37] propose *effective web governance* through the application of behavioral interventions to improve decision-making in a web context, e.g., by applying *nudges*. Nudges are interventions that subtly guide users to make better decisions without restricting possible choices, e.g., by setting defaults, creating friction and altering the required effort, or suggesting alternatives [8, 62]

Caraban et al. [8] grouped different nudging approaches according to their level of *transparency* (non-transparent, transparent) and *mode of thinking engaged* (automatic mind, reflective mind), following the categories proposed by [23] (see Figure 1). The distinction between automatic and reflective nudging approaches is closely related to the *Elaboration Likelihood Model* by Petty and Cacioppo [47]. The Elaboration Likelihood Model is a theoretical framework that distinguishes between the *peripheral* and the *central* route of processing persuasive interventions such as nudges. Automatic nudging, which operates through the peripheral route of processing, aims at *influencing behavior* by relying on simple, non-argumentative cues to evoke intuitive and unconscious

reactions. Reflective nudging, which operates through the central route of persuasion, aims at *prompting reflective choice* by engaging the critical thinking skills of the recipient to evaluate the arguments presented in a message.

The use of automatic nudges has received criticism for being paternalistic, harming user autonomy, decreasing user experience, hindering learning, and resulting in habituation effects [8, 23, 29]. Yet, purely reflective nudging approaches may not be suitable either in the context of bias mitigation. Processing reflective nudges could further increase cognitive demand and thus the susceptibility to cognitive biases.

Prior research on confirmation bias mitigation during web interactions with information items investigated interventions with different objectives: *facilitating information processing*, e.g., with data visualization [36] or argument summaries [55]; *increasing exposure to selected items*, e.g., with preference-inconsistent recommendations [56] or alternative query suggestions [51]; or *raising visibility of behavior*, e.g., with feedback on the political leaning of a user's reading behavior [43].

To mitigate confirmation bias during search result selection, interventions that aim at *decreasing exposure to selected items*, namely attitude-confirming search results, may also be effective. While such interventions have not yet been investigated for confirmation bias mitigation during web search, they have been researched in a different context—to prevent engagement with mis- and disinformation. A particularly successful approach that has been applied across different social networking platforms consists of *warning labels* to flag items that may contain misinformation and *obfuscations* to decrease the ease of access to these items by default [10, 29, 40]. Categorizing these interventions according to the taxonomy by Caraban et al. [8], they combine reflective and automatic nudging elements: They *prompt reflective choice* by confronting users with the risk of engaging with a given item through the warning label and *influence behavior* by decreasing the ease of access to the item through default obfuscations that can be removed with additional effort. Similar interventions that decrease exposure to attitude-confirming items could mitigate confirmation bias during search result selection.

2.3 Cognitive Reflection and Engagement with Information

Search behavior, susceptibility to cognitive biases, and reaction to nudging approaches are affected by various context-dependent user states and relatively stable user traits. A relatively stable user trait in the context of engagement with information is a user's *cognitive reflection* style. The concept is closely related to the *need for cognition*, an individual's tendency to organize their experience meaningfully [6, 15]. An individual's cognitive reflection style can be captured with the ***Cognitive Reflection Test*** **(CRT)** [15]. People with a high CRT score are considered to rely more on analytic thinking, thus enjoying challenging mental activities. People with a low CRT score, on the other hand, are considered to rely more on intuitive thinking, thus enjoying effortless information processing [6, 11, 15].

This general tendency of relying on either more analytic or intuitive thinking affects different aspects of engaging with information [7, 46, 64]. Searchers with an analytic cognitive style were observed to invest more cognitive effort in information search [67]. Compared to more intuitive thinkers, analytic individuals were further found to more effectively overcome uncertainties, critically assess their arguments, and monitor their thinking during learning tasks in an online environment [58]. Coutinho [12] found that a more analytic cognitive style is positively correlated with higher metacognitive skills, hence with increased thinking about thinking, a more accurate self-assessment, and increased awareness of one's behavior.

Users' cognitive reflection style was observed to impact whether and how users engage with false information and information that they perceive to be untrustworthy [42, 46, 64]. Tsfati and Cappella [64] observed that more analytic people are more likely than intuitive people to

engage with information from sources they do not perceive as trustworthy. The authors reason that analytic people do so because they want to make sense of the world and learn about different viewpoints while intuitive people tend to avoid exposure to mistrusted sources. Pennycook and Rand [46] found that analytic users more accurately detect fake news than intuitive users, even if the false information aligns with their ideology. Mosleh et al. [42] observed that intuitive users are generally more gullible (i.e., more likely to share money-making scams and get-rich schemes). They further observed cognitive echo chambers, emerging clusters of accounts of either analytic or intuitive social media users.

Whether people are generally more intuitive or analytic thinkers is a contributing factor to their susceptibility to peripheral (i.e., automatic nudging elements) or central (i.e., reflective nudging elements) cues of persuasion [7]. In the context of nudging, intuitive thinkers might thus be more inclined to follow automatic nudging and choose the path of lowest effort which leads to an unconscious change in their behavior. Analytic thinkers, on the other hand, might be more inclined to follow reflective nudging elements and actively decide to change their behavior.

3 WARNING LABEL AND OBFUSCATION STUDY

With the work presented in this article, we aim to understand the benefits and risks of an intervention to support unbiased search on debated topics. Therefore, with our first preregistered user study,[1] we tested the following hypothesis[2,3]:

H1: *Search engine users are less likely to click on attitude-confirming search results when some search results on the* **search engine result page***(***SERP***) are displayed with a warning label with obfuscation.*

We conducted a between-subjects user study to test this hypothesis. We manipulated the **search result display** (*targeted warning label with obfuscation, random warning label with obfuscation, regular*) and evaluated participants' **clicks on attitude-confirming search results**. To gain a more comprehensive understanding of the potential benefits and risks of this intervention on search behavior and searchers and uncover potential variations among individuals, we investigated trends in supplementary exploratory data that we collected with this user study. This exploratory data comprises participants' cognitive reflection style, their engagement with the warning label and obfuscated search results (**clicks on show-button**, **clicks on search results with warning labels**), as well as participants' reflection after the interaction (**attitude change**, **accuracy bias estimation**). Note that, throughout the article, all analyses labeled as *exploratory* were not preregistered.

3.1 Method

3.1.1 Experimental Setup. All related material, including the pre- and post-search questionnaires, can be found at the link in Footnote 1.

Topics and Search Results. The dataset contains search results for the following four debated topics: **(1)** Is Drinking Milk Healthy for Humans? **(2)** Is Homework Beneficial? **(3)** Should People Become Vegetarian? **(4)** Should Students Have to Wear School Uniforms? For each of these, viewpoint and relevance annotations were collected for 50 search results. Out of this dataset of 200 search results, 12 randomly selected search results with overall balanced viewpoints (two

[1]The preregistration of this study can be found in our repository: https://osf.io/32wym/?view_only=19cf6003ec1b45c29dbd537058d14b4f

[2]Next to H1, we tested additional hypotheses on the task design and behavioral patterns across tasks in this user study. The results are not relevant to the focus of this article. They can be found in Rieger et al. [54].

[3]We reformulated some research questions and hypotheses to ensure consistency in wording across both studies. In terms of content, they remain the same as in the preregistrations.

strongly supporting, two *supporting*, two *somewhat supporting*, two *somewhat opposing*, two *opposing*, and two *strongly opposing*) on one of the four topics were displayed to the participants.

Warning labels and Obfuscation. In the search result display conditions with intervention, results were obfuscated with a warning label, warning of the risk of confirmation bias and advising the participant to select another item (see (1) in Figure 2). The warning label included a link to the *Wikipedia* entry on confirmation bias [71] so that participants could inform themselves. To view the obfuscated search result, participants had to click a button, stating they were aware of the risk of confirmation bias.

Cognitive Reflection Test. We measured participants' cognitive style in the post-interaction questionnaire with the **CRT** [15]. To avoid an effect of familiarity with the three questions of this widely used test, we reworded the three questions in the following way:

(1) A toothbrush and toothpaste cost $2.50 in total. The toothbrush costs $2.00 more than the toothpaste. How much does the toothpaste cost? *intuitive: $0.50, correct: $0.25*
(2) If it takes 10 carpenters 10 hours to make 10 chairs, how many hours would it take 200 carpenters to make 200 chairs? *intuitive: 200 hours, correct: 10 hours*
(3) On a pig-farm cases of a pig-virus were found. Every day the number of infected pigs doubles. If it takes 28 days for the virus to infect all pigs on the farm, how many days would it take for the virus to infect half of all pigs on the farm? *intuitive: 14 days, correct: 27 days*

3.1.2 Procedure. The data was collected via the online survey platform *Qualtrics*.[4] The user study consisted of the three following steps:

(1) Pre-interaction questionnaire: Participants were given 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.* They were asked to state their attitude on the four topics on a seven-point Likert scale ranging from *strongly agree* to *strongly disagree* **(prior attitude)**. Subsequently, they were randomly assigned to one of the topics for which they reported to strongly agree or disagree. If they did not report to strongly agree or disagree on any topic, they were randomly assigned to one of the topics for which they reported to agree or disagree. If participants did not fulfill this requirement (i.e., reported weak attitudes on all topics), they were not able to participate further but received partial payment, proportional to the time invested in the task. For the assigned topic, they were asked to state their knowledge on a seven-point Likert scale ranging from *non-existent* to *excellent* **(self-reported prior knowledge)**.

(2) Interaction with the search results: Participants were randomly assigned to one of the three search result display conditions (*targeted warning label with obfuscation, random warning label with obfuscation, regular*) **(search result display)**. Moreover, they were assigned to one out of two task conditions, in which we asked participants to explore the search results by clicking on search results and retrieving the linked documents and mark search results that they considered to be particularly relevant and informative either simultaneously, or in two subsequent steps (for details see [54]). With this article, however, we focus exclusively on searchers' exploration (i.e., clicking) behavior. Since we did not find differences in clicking interactions between both task conditions, these conditions are combined into a single group for all subsequent analyses.

For the search task, participants were exposed to 12 viewpoint-balanced search results on their assigned topic. Of those, four search results were initially displayed with a warning label with obfuscation in the targeted and random warning label with obfuscation conditions. To reveal the obfuscated search results, participants could click on a button, from here on referred to as

[4]Qualtrics: https://www.qualtrics.com

show-button **(clicks on show-button)**. From the interaction logs, we calculated the proportion of participants' **clicks on attitude-confirming search results**. For participants in the targeted and random warning label with obfuscation conditions, we calculated the proportion of **clicks on search results with warning labels**. We did not include a time limit in either direction to enable natural search behavior (as far as this is possible in a controlled experimental setting). However, data of participants who did not click on any search result and/or who spent less than one minute exploring the SERP was excluded before data analysis.[5]

(3) Post-interaction questionnaire: Participants were asked to state their attitude again **(attitude change)**. Furthermore, they were asked to reflect and report on their search result exploration on a 7-point Likert scale ranging from *all search results I clicked on opposed my prior attitude* to *all search results I clicked on supported my prior attitude* **(accuracy bias estimation)**. To conclude the task, participants were asked to answer the three questions of the CRT **(cognitive reflection)**.

3.1.3 Variables.

- Independent Variable: **Search result display** (categorical). Participants were randomly assigned to one of three display conditions (see warning label and obfuscation study in Figure 3): (1) targeted warning label with obfuscation of extreme attitude-confirming search results, (2) random warning label with obfuscation of four randomly selected search results, and (3) regular (no intervention).
- Dependent Variable: **Clicks on attitude-confirming search results** (continuous). The proportion of attitude-confirming results among the search results participants clicked on during search results exploration.
- Exploratory Variables:
  - **Clicks on search results with warning labels** (continuous). For targeted and random warning label with obfuscation condition: Proportion of obfuscated results among the search results participants clicked on during search results exploration.
  - **Cognitive reflection** (categorical). Participants' cognitive reflection style was measured with an adapted version of the Cognitive Reflection Task (see Section 3.1) in the post-interaction questionnaire. Participants with zero or one correct response were categorized as **intuitive**, and participants with two or three correct responses were categorized as **analytic**.
  - **Clicks on show-button** (discrete). Number of clicks on unique *show-buttons* (up to 4) to reveal an obfuscated search result (only in conditions with obfuscation).
  - **Attitude change** (discrete). Difference between attitude reported on a seven-point Likert scale, ranging from *strongly disagree* (−3) to *strongly agree* (3) in the pre-interaction questionnaire and the post-interaction questionnaire. Attitude difference is encoded in a way that negative values signify a change in attitude toward the opposing direction, whereas positive values indicate a reinforcement of the attitude in the supportive direction. Since we only recruited participants with moderate and strong prior attitudes (−3, −2, 2, 3), the values of attitude change can range from −6 (change from +3 to −3, or −3 to +3) to 1 (change from +2 to +3, or −2 to −3).
  - **Accuracy bias estimation** (continuous). Difference between (a) *observed bias* (as the proportion of attitude-confirming clicks) and (b) *perceived bias* (reported in the post-interaction questionnaire and re-coded into values from 0 to 1). Values range from −1

[5]In a pre-test, we observed that participants who spent less than a minute engaged notably less with the search page. We thus applied the one-minute cut-off to filter out low-quality data from crowdworkers who satisficed by minimizing the amount of effort invested in the task [30].

Table 1. Distribution across Conditions in Warning Label and Obfuscation Study: Number of Participants Per Search Result Display Conditions and Topic

	Topic 1	Topic 2	Topic 3	Topic 4	All
targeted w + obf	20	32	19	31	**102**
random w + obf	23	27	19	31	**100**
regular	15	22	20	23	**80**
All	**58**	**81**	**58**	**85**	**282**

1: Is Drinking Milk Healthy for Humans?; 2: Is Homework Beneficial?; 3: Should People Become Vegetarian?; 4: Should Students Have to Wear School Uniforms?

to 1, with positive values indicating an overestimation and negative values and underestimation of bias.

— **Self-reported prior knowledge** (discrete). Reported on a seven-point Likert scale ranging from *non-existent* to *excellent* as a response to how they would describe their knowledge on the topic they were assigned to.

— **Usability and Usefulness** (continuous). Mean of responses on a seven-point Likert scale to the modules *usefulness, usability* (six items) from the *meCUE 2.0*[6] questionnaire.

To describe the sample of study participants, we further asked them to report their age and gender.

3.2 Results

3.2.1 Description of the Sample. An a priori power analysis for a between-subjects ANOVA (with $f = 0.25$, $\alpha = \frac{0.05}{4} = 0.0125$ (due to initially testing four different hypotheses, see Footnote 2), and (1- β) = 0.8) determined a required sample size of 282 participants. 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). To achieve the required sample size, we employed a staged recruitment approach, sequentially recruiting participants and monitoring the number of participants that fulfill the inclusion criteria detailed below. For that, we recruited a total of 510 participants via the online participant recruitment platform *Prolific*.[7] From these 510 participants, 228 were excluded from data analysis for failing the following preregistered inclusion criteria: They did not report having 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 any search results *(57)*. We paid all participants regardless of whether we excluded their data from the analysis.

Our final dataset consisted thus of 282 participants, of which 51% reported to be male, 49% female, <1% non-binary/other. Concerning the age of the participants, 49.6% reported to be between 18 and 25, 27.3% between 26 and 35, 12.1% between 36 and 45, 7.1% between 46 and 55, 3.5% between 56 and 65, and 0.4% more than 65 years old.

The task in each display condition was completed by 80 to 102 participants and 58 to 85 participants saw search results of the different topics (see Table 1). The mean time spent exploring the SERP was 4min 45sec ($SE = 15.6sec$), ranging from a minimum of 1 min to a maximum of 26 min, with no evidence for differences between search result display conditions ($F(2,279) = 0.34, p = .71, f = 0.05$). The mean number of clicks on search results was 3.26 ($SE = 0.13$), approximately 25% of the 12 displayed search results, with no evidence for differences between search result display conditions ($F(2,279) = 0.88, p = .42, f = 0.08$).

[6]meCUE usability scale: http://mecue.de/english/home.html
[7]Prolific: https://www.prolific.co/

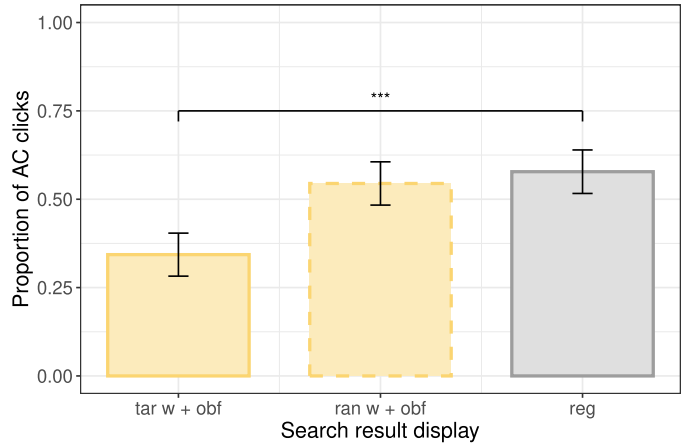

Fig. 4. Study 1: **Clicks on attitude-confirming search results**. Mean proportion of participants' attitude-confirming clicks per search result display condition (targeted warning label with obfuscation, random warning label with obfuscation, regular) with 95% confidence intervals. A proportion of one implies that all clicks were on attitude-confirming search results.

3.2.2 Hypothesis Testing: Effect of Search Result Display on Clicks on Attitude-Confirming Search Results. Although the distribution of attitude-confirming clicks did not exhibit normality, it is worth noting that ANOVAs have shown robustness in studies involving large sample sizes, even in cases where normality assumptions are not met [4, 72]. Considering this, we opted to employ ANOVAs for the statistical assessment of variations in participants' click behavior. The results of the ANOVA show evidence for a moderate effect of **search result display** on clicks on attitude-confirming search results ($F(2,279) = 17.14, p < .001, f = 0.35$).[8] A pairwise post hoc Tukey's test shows that the proportion of clicks on attitude-confirming search results was significantly lower for participants who were exposed to **targeted warning labels with obfuscations** ($mean = 0.34, SE = 0.03$) compared to those who saw **random warning labels with obfuscations** ($mean = 0.55, SE = 0.03; p < .001$), and those who saw **regular** search results ($mean = 0.58, SE = 0.03; p < .001$; see Figure 4). However, there was no evidence for a difference in the clicking behavior between **random warning labels with obfuscations** and **regular** search result display.

3.2.3 Exploratory Observations. We inspected the exploratory data to derive new hypotheses by visually investigating plots of means and standard errors, as well as boxplots of the (exploratory) dependent variables **clicks on search results with warning labels**, **clicks on show-button**, **attitude change**, and **accuracy bias estimation** for the (exploratory) independent variables **search result display** and **cognitive reflection**. We observed that participants who, according to the CRT, are more analytic thinkers were more likely to engage with search results with warning labels and to click on the show-button (see Figures 5 and 6). Furthermore, participants' attitude change seemed to be influenced by the display condition and their cognitive reflection style (see Figure 7). We also noted that participants who were exposed to targeted warning labels with obfuscations tended to overestimate their confirmation bias. Analytic participants more accurately estimated their bias than intuitive participants (see Figure 8).

[8]We validated the ANOVA results by additionally applying a Kruskal-Wallis test which likewise yielded a moderate effect ($H(2) = 33.87, p < .001, \eta^2 = 0.11$).

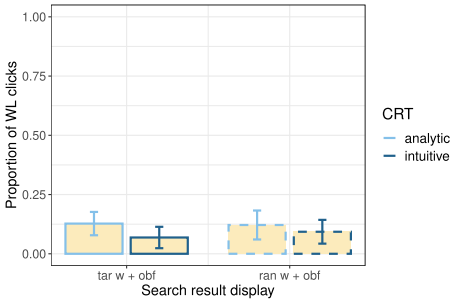

Fig. 5. Study 1 (exploratory): **Clicks on search results with warning labels**. Mean proportion of clicks on search results that were displayed with a warning label per search result display condition (targeted warning label with obfuscation, random warning label with obfuscation) and cognitive reflection style (analytic, intuitive) with 95% confidence intervals.

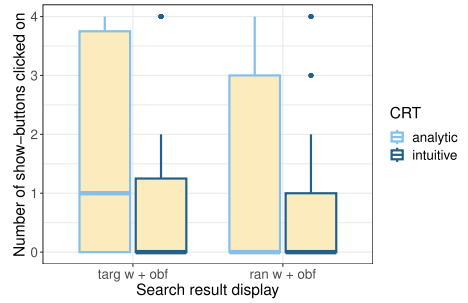

Fig. 6. Study 1 (exploratory): **Engagement with warning labels**. Boxplots with medians and quartiles, illustrating the distribution of the number of show-buttons that each participant clicked on (up to four) per search result display condition (targeted warning label with obfuscation, random warning label with obfuscation) and cognitive reflection style (analytic, intuitive).

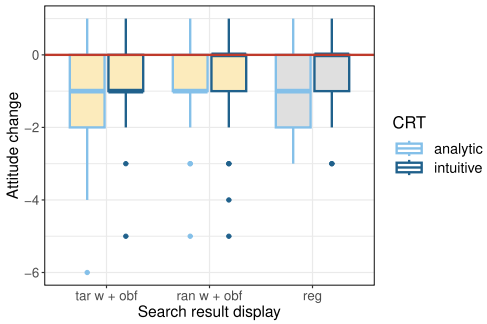

Fig. 7. Study 1 (exploratory): **Attitude change**. Boxplots with medians and quartiles, illustrating the distribution of participants' difference between pre- and post-interaction attitude per search result display condition (targeted warning label with obfuscation, random warning label with obfuscation, regular) and cognitive reflection style (analytic, intuitive). Negative values indicate a weakening of the initial attitude.

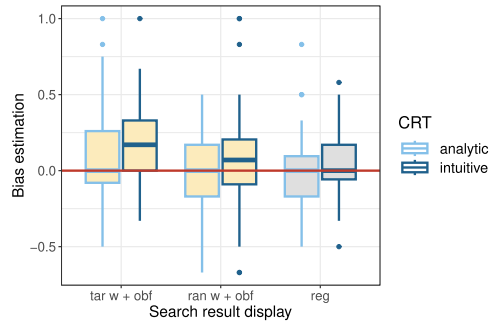

Fig. 8. Study 1 (exploratory): **Accuracy of bias estimation**. Boxplots with medians and quartiles, illustrating the distribution of participants' difference between observed bias and perceived bias per search result display condition (targeted warning label with obfuscation, random warning label with obfuscation, regular) and cognitive reflection style (analytic, intuitive). Positive values indicate an overestimation of bias (i.e., perceived bias is higher than observed bias in behavior).

We further explored means and standard errors of clicks on attitude-confirming search results across different degrees of self-reported **prior knowledge**, yet no differences emerged. Finally, we investigated whether participants in distinct **search result display** conditions exhibited different levels of usefulness and usability. The inspection of means and standard errors revealed no discernible differences between the three conditions (see Table 2).

3.3 Reflections and Follow-Up Hypotheses

We found that targeted obfuscations with warning labels decreased the likelihood of clicking on attitude-confirming search results. However, it is unclear whether the intervention prompted

Table 2. Study 1 (Exploratory): **Usability and Usefulness**

	Usability		Usefulness	
	mean	*SE*	*mean*	*SE*
Targeted w + obf	6.06	0.09	5.47	0.1
Random w + obf	6	0.1	5.52	0.11
Regular	6.24	0.11	5.59	0.11

Mean usability and usefulness scores with standard error per search result display condition (targeted warning label with obfuscation, random warning label with obfuscation, regular).

reflective choice, and participants read the warning label and clicked on the show-button to reveal the search result but, now aware of confirmation bias, actively decided to interact less with attitude confirming search results; or the intervention *automatically* influenced behavior, and participants engaged less with obfuscated items because interaction with those required additional effort.

Our exploratory findings indicate that both targeted and random warning labels decrease engagement with search results with warning labels and that intuitive searchers are less likely to engage with the warning label by clicking on the show-button than analytic searchers. This could imply that, in line with the Elaboration Likelihood Model [47], for more intuitive users, decreased engagement might be caused primarily by the obfuscation. Yet, if intuitive users do not engage with the intervention and ignore the warning label, the intervention might effectively not be transparent and manipulate instead of influence user behavior (see Figure 1).

To understand how different searchers are impacted by the *reflective* and *automatic* elements of the intervention, we need to investigate the effects of warning labels and obfuscations separately (**warning labels with and without obfuscations**). Based on our exploratory insights, we suggest the following primary hypotheses[3] for this follow-up study:

- **H2a:** Search engine users are less likely to click on search results that are displayed with a warning label with obfuscation than search results that are displayed with a warning label without obfuscation.
- **H2b:** Intuitive search engine users are less likely to click on a button to reveal an obfuscated search result than analytic users.
- **H2c:** The difference in clicks on search results that are displayed with a warning label without obfuscation compared to those with obfuscation is moderated by users' cognitive reflection style.
- **H2d:** Clicks on search results that are displayed with a warning label with obfuscation will be reduced, while clicks on search results with a warning label without obfuscation will only be reduced when they are applied to attitude-confirming search results (targeted) but not when they are applied incorrectly, to random search results.
- **H2e:** The moderating effect of targeting on the effect of warning style on users' clicks on search results with warning labels is moderated by users' cognitive reflection style.

Furthermore, based on our exploratory observations on attitude change and accuracy of bias estimation, we suggest the following secondary hypotheses[3]:

- **H3a:** Attitude change is greater in conditions with targeted warning labels than in conditions with random warning labels and no warning labels.
- **H3b:** The effect of the search result display condition on attitude change is moderated by participants' cognitive reflection style.

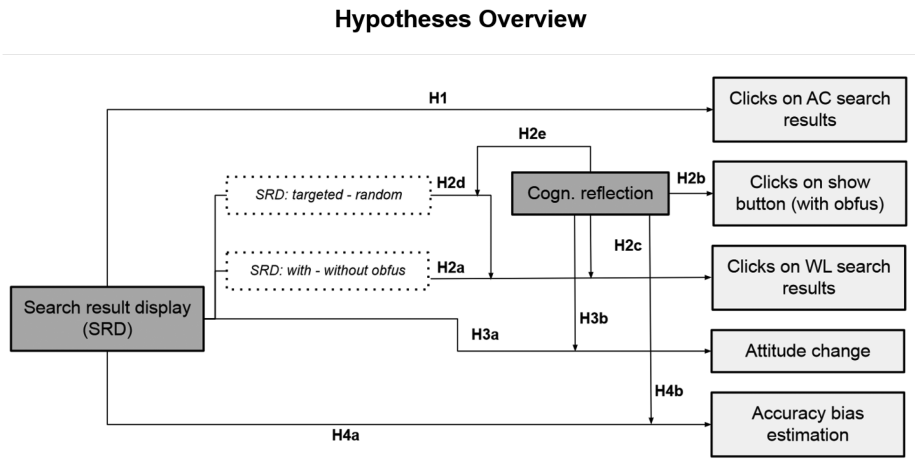

Fig. 9. Study 2: Automatic vs. reflective study. Overview of hypotheses with independent (dark gray) and dependent (light gray) variables.

— **H4a:** Users who see search results with targeted warning labels overestimate the confirmation bias in their clicking behavior to a greater extent than users who see search results with random or no warning labels.
— **H4b:** Analytic participants make more accurate estimations of the bias in their behavior while intuitive participants tend to overestimate the bias in their behavior.

4 FOLLOW-UP: AUTOMATIC VS. REFLECTIVE STUDY

We conducted a follow-up study, the **automatic vs. reflective study**, with the primary goal to better understand the effect of warning labels and obfuscations on different users' search behavior. Specifically, we investigated whether the observed effect was caused by the obfuscation (automatic) or the warning label (reflective) (H2a, H2d). With this follow-up study, we also tested whether we could replicate the findings we made in the **warning label and obfuscation study** for different search results, but the same topics (H1). To better understand the impact of the interventions on the searcher, we further tested whether the search result display has effects on their attitude change (H3a) and awareness of bias (H4a). Finally, we investigated the potential (moderating) effects of participants' tendency to be more intuitive or analytic thinkers, according to their CRT scores, on their engagement with the intervention (H2b), engagement with search results with warning labels (H2c, H2e), attitude change (H3b), and accuracy of bias estimation (H4b) (see Section 3.3 and Figure 9).

4.1 Method

The method we used for the second, preregistered,[9] between-subjects user study was essentially identical to the method we used for the first user study. We made the following minor changes to permit testing the follow-up hypotheses (H2–H4, see Section 3.3):

- **Search result display:** To allow us to understand the distinct impact of the *automatic* (obfuscation), and the **reflective** (warning label) nudging element of the intervention, we introduced two additional **search result display** conditions: targeted and random warning

[9]The preregistration of the second user study can be found in our repository: https://osf.io/p3ykv/?view_only=93f2eebbd55445aea3604ae751127892

label without obfuscation (see (2) in Figure 2). This resulted in the following five display conditions (see Figure 3):

(1) **targeted** warning label **with obfuscation** of moderate and extreme attitude confirming search results
(2) **targeted** warning label **without obfuscation** of moderate and extreme attitude confirming search results
(3) **random** warning label **with obfuscation** of four randomly selected search results
(4) **random** warning label **without obfuscation** of four randomly selected search results
(5) regular (no intervention)

- **Experimental Setup:** To test the reproducibility of the findings in the **warning label and obfuscation study** for different search results, we randomly sampled new search results (12 per topic, two *strongly supporting*, two *supporting*, two *somewhat supporting*, two *somewhat opposing*, two *opposing*, two *strongly opposing*) for the same topics from the set of viewpoint annotated search results which we collected for the **warning label and obfuscation study**. Since concerns about the validity of the CRT have been raised [21, 63], we included the exploratory variable of participants' *need for cognition*, a measure that captures users' motivation to engage in effortful thinking, to support potential findings on moderating effects of cognitive reflection. We captured participants' need for cognition with a self-report with a 4-item subset of the need for cognition questionnaire by Cacioppo et al. [6]. These four items include the same subset as used in Buçinca et al. [5]: *I would prefer complex to simple problems; I like to have the responsibility of handling a situation that requires a lot of thinking; Thinking is not my idea of fun; I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.*
- **Variables:** Exploratory variables in the **warning label and obfuscation study** were turned into independent and dependent variables in the **automatic vs. reflective study**. In the **automatic vs. reflective study**, we thus manipulated and measured the following variables:
  - **Independent Variables:** Search result display, cognitive reflection
  - **Dependent Variables:** Clicks on attitude-confirming search results (attitude-confirming), clicks on search results with warning labels, clicks on show-button, attitude change, accuracy bias estimation
  - **Exploratory Variables:** Need for cognition, prior knowledge, usability and usefulness
- **Procedure:** The procedure of data collection remained essentially the same as described in Section 3.1 for the **warning label and obfuscation study**. The four questions to capture need for cognition were added to the post-interaction questionnaire. We slightly increased the reward for participation to 1.80£ (mean = 7.89£/h) to adhere to the updated Prolific suggestion. Furthermore, we launched the data collection in multiple batches at different times of the day and night, to increase the likelihood of a sample with high diversity in geographical locations.
- **Attention checks:** To adhere to Prolific guidelines, we included an additional attention check, leading to a total of five, and adapted the exclusion criterion to failing two or more (instead of one or more out of four) attention checks.

4.2 Results

4.2.1 Description of the Sample. An a-priori power analysis for between-subjects ANOVAs, assuming moderate effects ($f = 0.25$, $\alpha = \frac{0.05}{10} = 0.005$ (due to testing 10 hypotheses), (1- β) = 0.8, up to 10 groups) determined a required sample size of 307 participants. As for the **warning label and obfuscation study**, we employed a staged recruitment approach in which we recruited an

Table 3. Distribution across Conditions in Automatic vs. Reflective Study: Number of Participants Per Search Result Display Conditions and Topic

	Topic 1	Topic 2	Topic 3	Topic 4	All
targeted w + obf	18	19	7	18	**62**
targeted w	18	16	14	19	**67**
random w + obf	11	25	11	15	**62**
random w	11	13	9	20	**53**
regular	25	20	5	13	**63**
All	**83**	**93**	**46**	**85**	**307**

1: Is Drinking Milk Healthy for Humans?; 2: Is Homework Beneficial?; 3: Should People Become Vegetarian?; 4: Should Students Have to Wear School Uniforms?

overall of 481 participants. Of these, 174 were excluded because they did not fulfill the inclusion criteria: They did not report having a strong attitude on any of the topics *(31)*, failed at two or more of five attention checks *(2)*, spent less than 60 seconds on the SERP*(88)*, or did not click on any search results (*53*). Of the 307 included participants, 52% reported to be male, 46% female, 2% non-binary/other, and <1% preferred not to share their gender. Furthermore, 40.7% reported to be between 18 and 25, 37.1% between 26 and 35, 12.7% between 36 and 45, 6.8% between 46 and 55, 1.6% between 56 and 65, and 1% more than 65 years old.

A total of 53–67 participants completed the task in each of the five search result display conditions and 46–93 participants saw search results for each of the four topics (see Table 3). Results of the CRT categorized 167 participants as *analytic* and 140 participants as *intuitive*. The mean time spent exploring the SERP page was 4 min 19 sec ($SE = 10.2sec$), ranging from a minimum of 1 min to a maximum of 19 min, with no evidence for differences between search result display conditions ($F(4,302) = 0.57, p = .69, f = 0.09$) and cognitive reflection categories ($F(4,302) = 2.18, p = .14, f = 0.08$). The mean number of clicks on search results was 2.8 ($SE = 0.09$), with no evidence for differences between search result display conditions ($F(4,302) = 1.24, p = .29, f = 0.13$), but a difference between cognitive reflection categories ($F(4,302) = 18.09, p < .001, f = 0.24$) with more clicks by analytic ($mean = 3.16, SE = 0.14$) than intuitive ($mean = 2.38, SE = 0.12$) participants.

4.2.2 Hypothesis Testing. We conducted five ANOVAs to test the 10 hypotheses and set the significance threshold at $\alpha = \frac{0.05}{10} = 0.005$, aiming at a type 1 error probability of $\alpha = 0.05$ and applying Bonferroni correction to correct for multiple testing.

H1: Main effect of search result display on attitude-confirming clicks (Replication). We could replicate the findings made in the **warning label and obfuscation study** by finding more evidence for a moderate effect of the **search result display** on **clicks on attitude-confirming search results** ($F(4,302) = 6.67, p < .001, f = 0.30$). A pairwise post hoc Tukey's test shows that the proportion of clicks on attitude-confirming search results was significantly lower for participants who were exposed to targeted warning labels with obfuscations ($mean = 0.34, SE = 0.03$) than those who were exposed to a regular search page ($mean = 0.53, SE = 0.04; p = .004$; see Figure 10). In comparison to the regular search page, participants exposed to targeted warning labels without obfuscations likewise exhibited a lower mean proportion of clicks on attitude-confirming search results ($mean = 0.41, SE = 0.03$). As in the **warning label and obfuscation study**, we did not observe lower proportions of clicks on attitude-confirming search results for participants exposed to random warning labels with obfuscations ($mean = 0.56, SE = 0.05$).

H2a: Main effect of obfuscation on clicks on search results with warning labels. We found evidence for a moderate effect of **obfuscation** on the proportion of clicks on search results that

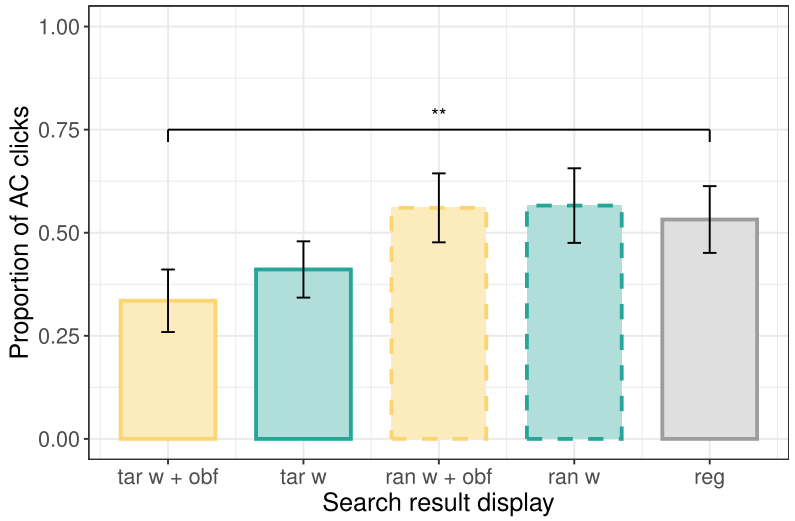

Fig. 10. Study 2: **Clicks on attitude-confirming search results**. Mean proportion of participants' attitude-confirming clicks per search result display condition (targeted warning label with obfuscation, targeted warning label without obfuscation, random warning label with obfuscation, random warning label without obfuscation, regular) with 95% confidence intervals. A proportion of one implies that all clicks were on attitude-confirming search results.

Table 4. Study 2 (Exploratory): **No Engagement with Warning Labels**

	CRT: analytic	CRT: intuitive	All
targeted w + obf	37%	63%	**48%**
random w + obf	58%	72%	**65%**
All	**47%**	**68%**	

The proportion of participants who did not engage with any warning label by clicking on the show-button per search result display condition (targeted warning label with obfuscation, random warning label with obfuscation) and cognitive reflection style (analytic, intuitive).

were displayed with a warning label ($F(1,236) = 12.9, p < .001, f = 0.23$). A post hoc Tukey test revealed that in conditions with obfuscations, participants clicked on fewer search results that were displayed with a warning label ($mean = 0.12, SE = 0.02$) than in conditions without obfuscations ($mean = 0.24, SE = 0.03; p < .001$; see Figure 11). Thus, H2a was confirmed.

H2b: Main effect of cognitive reflection on clicks of show-button. Descriptive statistics indicated that participants with an analytic as opposed to an intuitive **cognitive reflection** style were more likely to click on the show-button to reveal search results that were initially obfuscated (see Figure 12). However, evidence for this relation did not meet the Bonferroni-corrected significance threshold of $\alpha = 0.005$ ($F(1,122) = 6.22, p = .014, f = 0.23$). To gain further insights, we explored (i.e., this analysis was not preregistered) the proportion of participants that did not at all engage with the warning label by clicking on the show-button and observed that overall, a high proportion of participants did not even once click on the show-button (56%). This exploratory analysis further revealed that more intuitive (68%) than analytic (47%) participants, and more participants in the random warning label condition (65%) than in the targeted warning label condition (48%) ignored the warning labels (see Table 4).

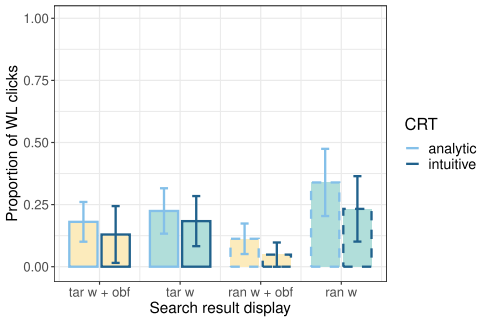

Fig. 11. Study 2: **Clicks on search results with warning labels**. Mean proportion of clicks on search results that were displayed with a warning label per search result display condition (targeted warning label with obfuscation, targeted warning label without obfuscation, random warning label with obfuscation, random warning label without obfuscation) and cognitive reflection style (analytic, intuitive) with 95% confidence intervals.

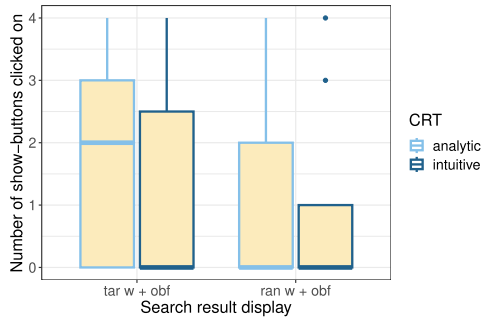

Fig. 12. Study 2: **Engagement with warning labels** (only for display conditions with obfuscation). Boxplots with medians and quartiles, illustrating the distribution of the number of show-buttons that each participant clicked on (up to four) per search result display condition (targeted warning label with obfuscation, random warning label with obfuscation) and cognitive reflection style (analytic, intuitive).

H2c: Interaction effect of cognitive reflection and obfuscation on clicks on search results with warning labels. We did not find evidence for an interaction effect of **cognitive reflection** and **obfuscation** on the proportion of clicks on search results that were displayed with a warning label ($F(1{,}236) = 0.04, p = .85, f = 0.01$; see Figure 11).

H2d: Interaction effect of targeting and obfuscation on clicks on search results with warning labels. Descriptive statistics suggest a disparity of the mean proportion of clicks on search results with warning labels between the conditions with and without obfuscations. This disparity was more pronounced in the random than in the targeted warning labels condition (see Figure 11). Yet, the interaction between **targeting** and **obfuscation** did not meet the Bonferroni-corrected significance threshold of $\alpha = 0.005$ ($F(1{,}236) = 5.41, p = .02, f = 0.15$).

H2e: Interaction effect of cognitive reflection, targeting, and obfuscation on clicks on search results with warning labels. We did not find evidence for an interaction effect of **cognitive reflection**, **targeting**, and **obfuscation** on the proportion of clicks on search results that were displayed with a warning label ($F(1{,}236) = 0.15, p = .70, f = 0.03$; see Figure 11).

H3a: Main effect of search result display on attitude change. We did not find evidence for an effect of **search result display** on participants' **attitude change** ($F(4{,}297) = 1.55, p = .18, f = 0.14$; see Figure 13).

H3b: Interaction effect of cognitive reflection and search result display on attitude change. We did not find evidence for an interaction of **cognitive reflection** and **search result display** on **attitude change** did not meet the Bonferroni-corrected significance threshold of $\alpha = 0.005$ ($F(4{,}297) = 2.72, p = .03, f = 0.19$); see Figure 13).

H4a: Main effect of search result display on accuracy of bias estimation. We did not find evidence for an effect of **search result display** on participants' **accuracy of bias estimation** ($F(4{,}297) = 0.77, p = .55, f = 0.10$; see Figure 14).

H4b: Interaction effect of cognitive reflection and search result display on accuracy of bias estimation. We did not find evidence for an interaction effect of **cognitive reflection** and **search result display** on participants' **accuracy of bias estimation** ($F(4{,}297) = 0.62, p = .64, f = 0.09$; see Figure 14).

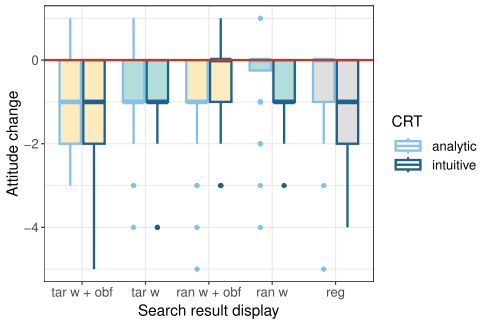

Fig. 13. Study 2: **Attitude change**. Boxplots with medians and quartiles, illustrating the distribution of participants' difference between pre- and post-interaction attitude per search result display condition (targeted warning label with obfuscation, targeted warning label without obfuscation, random warning label with obfuscation, random warning label without obfuscation, regular) and cognitive reflection style (analytic, intuitive). Negative values indicate a weakening of the initial attitude.

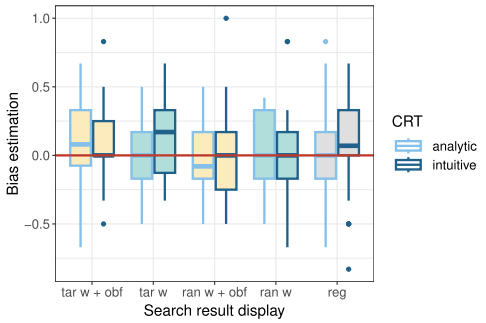

Fig. 14. Study 2: **Accuracy of bias estimation**. Boxplots with medians and quartiles, illustrating the distribution of participants' difference between observed bias and perceived bias per search result display condition (targeted warning label with obfuscation, targeted warning label without obfuscation, random warning label with obfuscation, random warning label without obfuscation, regular) and cognitive reflection style (analytic, intuitive). Positive values indicate an overestimation of bias (i.e., perceived bias is higher than observed bias in behavior).

4.2.3 Exploratory Observations. To gain deeper insights and support our findings from hypotheses testing, we explored the correlation between CRT and need for cognition, the potential effects of self-reported prior knowledge on engagement behavior and search consequences, and potential differences in usability and usefulness of the different search result display conditions for searchers with an analytic or intuitive cognitive reflection style. We calculated the *Spearman's* correlation coefficient between participants' CRT (behavioral) and need for cognition (questionnaire) score and found a weak positive relationship between the variables ($r = 0.21$, $p < .001$). Furthermore, we did not observe differences in any of the dependent variables between participants who reported a high compared to a low level of self-reported *prior knowledge*. Lastly, we did not observe any differences in questionnaire-reported usefulness and usability between the five *search result display* conditions. However, there was a tendency of participants who were categorized as *analytic* according to their CRT results to report lower usefulness of the SERP with targeted warning labels with and without obfuscations than participants who were categorized as *intuitive* (see Table 5). For the random and regular search result display conditions, no such difference was observed.

Participants who did not click on search results. The high rate of participants who had to be excluded from hypotheses testing because they did not click any search results ($N = 104$) prompted us to investigate possible causes. Our exploration revealed that there were no discernible differences in prior attitude strength or cognitive reflection style between the participants who clicked on search results and those who did not. Furthermore, the results indicate that participants who did not click on any search results were just as likely to change their attitude ($mean = -1.01, SE = 0.12$) as those who did click on one or more search results ($mean = -0.84, SE = 0.06$).

5 DISCUSSION

The two pre-registered user studies contribute to the understanding of behavioral interventions to support thorough and unbiased information-seeking strategies that are required for responsible

Table 5. Study 2 (Exploratory): **Usability and Usefulness**

	Usability				Usefulness			
	CRT: analytic		CRT: intuitive		CRT: analytic		CRT: intuitive	
	mean	*SE*	*mean*	*SE*	*mean*	*SE*	*mean*	*SE*
Targeted w + obf	5.75	0.18	6.14	0.16	5.43	0.17	6.17	0.17
Targeted w	5.99	0.12	6.01	0.19	5.51	0.18	5.96	0.15
Random w + obf	6.14	0.12	6.29	0.1	5.88	0.13	6.09	0.14
Random w	6.04	0.17	5.89	0.17	5.81	0.17	5.74	0.16
Regular	6.21	0.13	6.29	0.09	5.98	0.15	6	0.18

Mean usability and usefulness scores with standard error per search result display condition (targeted warning label with obfuscation, targeted warning label without obfuscation, random warning label with obfuscation, random warning label without obfuscation, regular) and cognitive reflection style (analytic, intuitive).

opinion formation on debated topics. Specifically, we focused on mitigating confirmation bias during search result selection by reducing engagement with attitude-confirming search results. Inspired by interventions to reduce engagement with misinformation, we applied warning labels and obfuscations to attitude-confirming search results. We further investigated the risks of the interventions by including conditions in which they were applied incorrectly, to random instead of attitude-confirming search results. To gain more comprehensive insights into potential effects of the interventions, we did not only investigate participants' search behavior, but additionally their attitude change and awareness of bias. We further investigated potential moderating effects of participants' cognitive reflection style. The following paragraphs summarise and discuss the findings and observations from both studies. Based on these findings, we discuss implications for designing interventions that aim at supporting thorough and unbiased information-seeking strategies.

5.1 Findings and Observations

5.1.1 Warning Label and Obfuscation. In the **warning label and obfuscation study**, we found that the intervention effectively reduced engagement. However, it reduced engagement with all search results that it was applied to, even if it was applied incorrectly to search results that were not attitude-confirming. This suggests that the intervention could be misused to manipulate engagement with information for alternative purposes, raising substantial ethical concerns.

The experimental setup did not allow for conclusions on how much of the effect was caused by the warning label (reflective element) versus the obfuscation (automatic element). To investigate potential effects of both nudging elements separately, we conducted a follow-up study and added a second intervention: We exposed participants to **warning labels without obfuscation** (see (2) in Figure 2).

5.1.2 Automatic vs. Reflective. We tested two interventions in the **automatic vs. reflective study**: warning label with obfuscation (reflective and automatic) and warning label without obfuscation (reflective). As before, we tested the interventions on either targeted attitude-confirming or random search results.

The mean proportion of clicks on attitude-confirming search results was reduced by targeted warning labels with and without obfuscations. This indicates that the mere warning label, thus the reflective element of the initial intervention, successfully achieves a reduction of clicks on attitude-confirming search results and thus mitigates confirmation bias. Thus, contrary to our concerns, the purely reflective intervention did not exhaust users' processing capacities.

The warning label alone, as opposed to with obfuscations, did not reduce clicks when they were applied incorrectly to random search results. Therefore, it seems that the automatic element is the

reason why searchers fail to detect and react to incorrect applications. These findings suggest that obfuscation restricts agency and harms autonomy. This is further supported by the high proportion of participants who seemed to have ignored the warning labels since they did not click on any show-button. While the intervention was designed with the intention to transparently influence behavior and prompt reflective choice, it might effectively manipulate behavior for users who do not engage with it.

These findings are in line with observations that users approach web search on debated topics with the intention to engage with diverse viewpoints [1, 39] but often fail to do so. For instance, [60] discuss that users have learned to trust that the resources provided by search engines, especially highly ranked results, are accurate and reliable. The authors reason that this might cause them to exert less cognitive effort in the search process. Yet, for complex search tasks that affect opinion formation, cognitive effort to engage with, compare, and evaluate different viewpoints would be required to form opinions responsibly [41]. Thus, interventions should encourage users to invest more effort into the search process to achieve their intended behavior of engaging with diverse viewpoints.

5.1.3 Cognitive Reflection Style. According to the Elaboration Likelihood Model [47], analytic thinkers might be more likely to follow reflective nudging elements, while intuitive thinkers might be more likely to follow automatic nudging elements. Thus, we investigated potential moderating effects of participants' cognitive reflection style on their engagement behavior.

In the **automatic vs. reflective study**, we did not find evidence for significant differences in engagement with the search results and interventions between users who, according to their CRT scores, are more analytic or intuitive thinkers. However, we did observe that, in line with the Elaboration Likelihood Model [47], the proportion of participants who did not at all engage with the warning labels is higher for intuitive (68%) than for analytic (47%) thinkers.

We attribute lack of evidence for a moderating effect of cognitive reflection style on clicks on the show-button on a combination of high noise in our data and strictly Bonferroni-corrected significance thresholds. The noise might have been caused by other user and context factors, such as their prior knowledge, situational and motivational influences (e.g., metacognitive states or traits), and ranking effects. Future research should thus continue to investigate the potential effects of users' cognitive reflection style and other user traits, states, and context factors that might moderate the effects of automatic and reflective elements of a nudge.

5.1.4 Attitude Change and Awareness of Bias. To gain more comprehensive insights into the potential effects of the intervention, we compared users' attitude change and awareness of bias between the different search result display conditions and cognitive reflection styles. We neither found evidence for differences between search result display conditions in participants' attitude change and awareness of bias nor for moderating effects of participants' cognitive reflection style. For both variables, we observed high levels of noise that might be caused by user differences beyond their cognitive reflection style.

In terms of responsible opinion formation, participants' prior knowledge of the topic should have a great impact on their attitude change. Users who have well-rounded prior knowledge should be less likely to change their attitude since it was already formed responsibly. Thus, it is unclear whether and what direction of attitude change would indicate responsible opinion formation.

Regarding awareness of bias, relatively stable traits and context-dependent states of users' metacognition (i.e., thinking about one's thinking) would likely have an impact and might have caused some of the observed noise. Of particular interest for responsible opinion formation and the risk of confirmation bias is users' *intellectual humility*, their ability to recognize the fallibility of their beliefs, and the limits of their knowledge [13, 49, 53]. Compared to people with low

intellectual humility, those with high intellectual humility were observed to invest more effort in information-seeking, spend more time engaging with attitude-opposing arguments [34, 50], and more accurately recognize the strength of different arguments, regardless of their stance [35]. Thus, high intellectual humility appears to reduce the likelihood of behavioral patterns that are common for confirmation bias [53]. The effect of metacognitive traits and states on search behavior and responsible opinion formation should be investigated in future research.

5.2 Implications

The observations and considerations discussed in the previous sections illustrate the complexity of researching and supporting web search for responsible opinion formation. The intervention of warning labels with obfuscations was inspired by approaches to combat misinformation. While we investigated this intervention because some objectives of combating misinformation overlap with those of mitigating confirmation bias during search, the research process and findings made us aware of a fundamental difference between them. Misinformation is a user-external threat and user behavior that is desired by system designers is fairly clearly defined (reduced/no engagement with items that contain misinformation). This is not the case for cognitive biases that impact search for opinion formation, which are user-internal and, depending on the context, serve a function [19].

As interventions to combat misinformation, the interventions we tested primarily aimed at reducing engagement with selected information items. To mitigate confirmation bias during search result selection, we aimed at reducing engagement with attitude-confirming search results. However, it is unclear what proportion of engagement with different viewpoints is desirable to support responsible opinion formation. When wanting to support users' in gaining a well-rounded knowledge, the desirable proportion likely depends on users' prior knowledge of the arguments for the different viewpoints. This illustrates that what constitutes *beneficial behavior* for responsible opinion formation during search on debated topics is non-trivial to define due to complex context and user dependencies.

Aiming for interventions that decide which information should be engaged with on the users' behalf imposes an immense level of responsibility on authorities who design them and decide on the application criteria [3]. Such interventions harm user autonomy and provide the means for abuse with intentions of stirring user behavior with (malicious) interests that do not align with the user's own interests. In preparation of our studies, we justified these risks of applying an automatic nudging element with the aim of reducing users' cognitive processing load. In fact, however, this was not necessary since users did not need the obfuscation, but chose to engage less with attitude-confirming search results when prompted to do so by a warning label without obfuscation. Thus, we may be underestimating users' abilities to actively choose unbiased behavior. Therefore, the risks of applying automatic nudging elements to support thorough information-seeking strategies are likely unwarranted. This potentially applies to other nudging scenarios in which the desired behavior is not clearly defined but depends on various (unknown) context and user factors.

Design Guidelines for Interventions. Given the complexity and potential far-reaching impact of search for opinion formation, we argue that interventions to support thorough and unbiased search should strictly emphasize user agency and autonomy. As a practical consequence, nudging interventions should prioritize reflective and transparent elements.

As an alternative to nudging interventions that steer user behavior directly, encouraging thorough information-seeking strategies could also be achieved by educating and empowering users to actively choose to change their behavior [53]. This can be done with boosting interventions that attempt to teach users to become resistant to various pitfalls of web interactions and remain effective for some time after being exposed to the intervention [25, 37]. Such approaches would improve user autonomy, minimize the risk of abuse and errors, and tackle the factors that impede

search for responsible opinion formation more comprehensively and sustainably [18, 25, 32, 37, 53]. Next to boosting, thorough information-seeking strategies that entail exploring, comparing, or evaluating different resources for sense-making and learning could be supported by other means of designing the search environment (e.g., adding metadata, such as stance labels) [59, 60].

Whether nudging, boosting, or other approaches, interventions that aim at supporting search for responsible opinion formation should be designed to increase transparency to and choice for the user [74]. This claim aligns with the EU's ethics guideline for trustworthy AI, which places human autonomy and agency at its core and states that AI systems (e.g., search engines) should support humans to make informed decisions by augmenting and complementing human cognitive skills instead of manipulating or herding them [27].

5.3 Limitations and Future Work

We acknowledge some limitations, mainly resulting from the controlled setting of this user study. We chose the controlled setting to be able to clearly distinguish the effects of the interventions from other factors that might affect search behavior. For that, we constructed an artificial scenario with one specific search task. Furthermore, we presented one specific set of pre-selected topics and viewpoint-labeled search results on a single SERP. While our objective was to closely assimilate real-world search settings, this controlled experimental setup did not allow participants to issue multiple queries or have access to great amounts of resources over an extended time period. Furthermore, while assigning participants to a topic for which they reported a strong attitude, we did not capture whether they were interested in learning about it. Future research should investigate whether the effects we observed will also be observed in less controlled search settings, how they evolve when users are exposed to the interventions for multiple search sessions, and whether the effects of the intervention are different for searchers who report weak prior attitudes on the topics.

We further attempted to ensure that ranking effects (i.e., position bias that causes more engagement with high-ranked items [20, 28]) would not distort the effects of the search result display by fully randomizing the ranking. Yet, given these known strong effects of search result ranking on user engagement, this design decision might have added noise to our data that prevented us from finding significant evidence for some of our hypotheses. Future work should thus investigate the interplay of interventions with ranking effects during search on debated topics.

Our representation of prior knowledge was limited. We did anticipate that prior knowledge could affect users' search behavior [16, 61] and attitude change, especially for users with strong opinions on debated topics. We thus captured users' self-reported prior knowledge. However, we did not find any effects of self-reported prior knowledge on user behavior, their attitude change, and the accuracy of bias estimation. Yet, this might be due to the low reliability of self-reported measures. Different levels of actual prior knowledge that we did not capture might have added further noise to our data. The effect of prior knowledge on search behavior, consequences, and metacognitive reflections during search for opinion formation should be investigated in future research.

Lastly, we investigated different factors of user engagement that might be impacted by the interventions, such as their clicking behavior, awareness of bias, and attitude change. However, we did not investigate additional variables that could indicate whether participants thoroughly explored the results (i.e., maximum scroll depth, dwell time), or whether they understood the encountered information (i.e., knowledge gain) and critically evaluated its arguments to form their opinion. Our explorations of data from participants who did not click on any search results revealed that those participants were just as likely to change their attitude. This observation indicates that the engagement variables captured in these user studies are not sufficient to model search consequences on learning and opinion formation. Future research should investigate searchers' engagement and

how it impacts learning and opinion formation more thoroughly, presumably by utilizing both quantitative and qualitative methods.

6 CONCLUSION

We conducted two user studies with the objective of understanding the benefits and risks of behavioral interventions to mitigate users' confirmation bias and support thorough and unbiased information-seeking strategies during search on debated topics. The findings from these studies indicate that obfuscations may risk manipulating behavior rather than guiding it while warning labels without obfuscations effectively encourage users to reduce their interaction with attitude-confirming search results. This suggests that when opting for automatic nudges to decrease cognitive load, users' capacity to actively choose unbiased behavior might be underestimated. We posit that ensuring and facilitating user agency is crucial for interventions that aim at supporting thorough and unbiased information behavior and that in cases where reflective nudging alternatives effectively encourage behavioral change, the risks associated with automatic nudges would not be justified. Obfuscations, and potentially other automatic nudging elements to guide search behavior, should thus be avoided. Instead, priority should be given to interventions that aim at strengthening human cognitive skills and agency, such as prompting reflective choice to engage with diverse viewpoints. This likely applies beyond our study context, extending to other nudging scenarios that can carry substantial consequences for individuals or society, in which determining what constitutes *beneficial behavior* (i.e., the target behavior toward which users should be nudged) is non-trivial due to complex context and user dependencies.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their helpful comments and suggestions to improve the article.

REFERENCES

[1] Marwah Alaofi, Luke Gallagher, Dana Mckay, Lauren L. Saling, Mark Sanderson, Falk Scholer, Damiano Spina, and Ryen W. White. 2022. Where do queries come from?. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, 2850–2862. DOI: https://doi.org/10.1145/3477495.3531711

[2] 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, 27–37. DOI: https://doi.org/10.1145/3406522.3446023

[3] Abraham Bernstein, Claes H. de Vreese, Natali Helberger, Wolfgang Schulz, and Katharina A. Zweig. 2020. Diversity, fairness, and data-driven personalization in (news) recommender system (dagstuhl perspectives workshop 19482). *Dagstuhl Manifestos* 9, 11 (2020), 117–124. DOI: https://doi.org/10.4230/DagRep.9.11.117

[4] M. José Blanca Mena, Rafael Alarcón Postigo, Jaume Arnau Gras, Roser Bono Cabré, and Rebecca Bendayan. 2017. Non-normal data: Is ANOVA still a valid option? *Psicothema* 29, 4 (2017), 552–557.

[5] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW 1 (2021), 188:1–188:21. DOI: https://doi.org/10.1145/3449287

[6] John T. Cacioppo, Richard E. Petty, and Chuan Feng Kao. 1984. The efficient assessment of need for cognition. *Journal of Personality Assessment* 48, 3 (1984), 306–307. DOI: https://doi.org/10.1207/s15327752jpa4803_13

[7] 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, 4 (1983), 805–818. DOI: https://doi.org/10.1037/0022-3514.45.4.805

[8] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–15.

[9] Noel Carroll. 2014. In search we trust: Exploring how search engines are shaping society. *International Journal of Knowledge Society Research* 5, 1 (2014), 12–27. DOI: https://doi.org/10.4018/ijksr.2014010102

[10] Sijing Chen, Lu Xiao, and Akit Kumar. 2023. Spread of misinformation on social media: What contributes to it and how to combat it. *Computers in Human Behavior* 141 (2023), 107643. DOI: https://doi.org/10.1016/j.chb.2022.107643

[11] 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, 2 (1955), 291–294. DOI: https://doi.org/10.1037/h0042761

[12] Savia A. Coutinho. 2006. The relationship between the need for cognition, metacognition, and intellectual task performance. *Educational Research and Reviews* 1, 5 (2006), 162–164. DOI: https://doi.org/10.5897/ERR.9000373

[13] Don E. Davis, Kenneth Rice, Stacey McElroy, Cirleen DeBlaere, Elise Choe, Daryl R. Van Tongeren, and Joshua N. Hook. 2016. Distinguishing intellectual humility and general humility. *The Journal of Positive Psychology* 11, 3 (2016), 215–224. DOI: https://doi.org/10.1080/17439760.2015.1048818

[14] Robert Epstein and Ronald E. Robertson. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences* 112, 33 (Aug. 2015), E4512–E4521. DOI: https://doi.org/10.1073/pnas.1419828112

[15] Shane Frederick. 2005. Cognitive reflection and decision making. *Journal of Economic Perspectives* 19, 4 (Nov. 2005), 25–42. DOI: https://doi.org/10.1257/089533005775196732

[16] Ujwal Gadiraju, Ran Yu, Stefan Dietze, and Peter Holtz. 2018. Analyzing knowledge gain of users in informational search sessions on the web. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. ACM, New York, NY, 2–11. DOI: https://doi.org/10.1145/3176349.3176381

[17] Lisa Gevelber. 2018. *It's All about 'me'—How People are Taking Search Personally*. Technical Report. Retrieved from https://www.thinkwithgoogle.com/marketing-strategies/search/personal-needs-search-trends/

[18] Gerd Gigerenzer. 2015. On the supposed evidence for libertarian paternalism. *Review of Philosophy and Psychology* 6, 3 (2015), 361–383. DOI: https://doi.org/10.1007/s13164-015-0248-1

[19] Gerd Gigerenzer and Henry Brighton. 2009. Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science* 1, 1 (2009), 107–143.

[20] Zhiwei Guan and Edward Cutrell. 2007. An eye tracking study of the effect of target rank on web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 417–420. DOI: https://doi.org/10.1145/1240624.1240691

[21] Matthew Haigh. 2016. Has the standard cognitive reflection test become a victim of its own success? *Advances in Cognitive Psychology* 12, 3 (2016), 145. DOI: https://doi.org/10.5709/acp-0193-5

[22] Alexander Halavais. 2017. *Search Engine Society*. John Wiley & Sons.

[23] Pelle Guldborg Hansen and Andreas Maaløe Jespersen. 2013. Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy. *European Journal of Risk Regulation* 4, 1 (2013), 3–28. DOI: https://doi.org/10.1017/S1867299X00002762

[24] Natali Helberger, Katharina Kleinen-von Königslöw, and Rob van der Noll. 2015. Regulating the new information intermediaries as gatekeepers of information diversity. *info* 17, 6 (2015), 50–71. DOI: https://doi.org/10.1108/info-05-2015-0034

[25] Ralph Hertwig and Till Grüne-Yanoff. 2017. Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science* 12, 6 (2017), 973–986. DOI: https://doi.org/10.1177/1745691617702496

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

[27] High-Level Expert Group on Artificial Intelligence. 2019. Ethics guidelines for trustworthy AI. European Commission. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[28] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately interpreting clickthrough data as implicit feedback. *ACM SIGIR Forum* 51, 1 (2017), 4–11. DOI: https://doi.org/10.1145/3130332.3130334

[29] Ben Kaiser, Jerry Wei, Eli Lucherini, Kevin Lee, J. Nathan Matias, and Jonathan Mayer. 2021. Adapting security warnings to counter online disinformation. In *Proceedings of the 30th USENIX Security Symposium*. 1163–1180.

[30] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. ACM, New York, NY, 1301–1318. DOI: https://doi.org/10.1145/2441776.2441923

[31] Hilary Kornblith. 1983. Justified belief and epistemically responsible action. *The Philosophical Review* 92, 1 (1983), 33–48. DOI: https://doi.org/10.2307/2184520

[32] Anastasia Kozyreva, Stephan Lewandowsky, and Ralph Hertwig. 2020. Citizens versus the internet: Confronting digital challenges with cognitive tools. *Psychological Science in the Public Interest* 21, 3 (2020), 103–156. DOI: https://doi.org/10.1177/1529100620946707

[33] Anastasia Kozyreva, Philipp Lorenz-Spreen, Stefan M. Herzog, Ullrich K. H. Ecker, Stephan Lewandowsky, Ralph Hertwig, Ayesha Ali, Joseph B. Bak-Coleman, Sarit Barzilai, Melisa Basol, Adam Berinsky, Cornelia Betsch, John Cook, Lisa Fazio, Michael Geers, Andrew M. Guess, Haifeng Huang, Horacio Larreguy, Rakoen Maertens, Folco Panizza, Gordon Pennycook, David G. Rand, Steve Rathje, Jason Reifler, Philipp Schmid, Mark Smith, Briony Swire-Thompson,

Paula Szewach, Sander van der Linden, and Sam Wineburg. 2022. Toolbox of interventions against online misinformation. DOI: https://doi.org/10.31234/osf.io/x8ejt

[34] Elizabeth J. Krumrei-Mancuso, Megan C. Haggard, Jordan P. LaBouff, and Wade C. Rowatt. 2020. Links between intellectual humility and acquiring knowledge. *The Journal of Positive Psychology* 15, 2 (2020), 155–170. DOI: https://doi.org/10.1080/17439760.2019.1579359

[35] Mark R. Leary, Kate J. Diebels, Erin K. Davisson, Katrina P. Jongman-Sereno, Jennifer C. Isherwood, Kaitlin T. Raimi, Samantha A. Deffler, and Rick H. Hoyle. 2017. Cognitive and interpersonal features of intellectual humility. *Personality and Social Psychology Bulletin* 43, 6 (2017), 793–813. DOI: https://doi.org/10.1177/0146167217697695

[36] 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*. ACM, New York, NY, 184–196. DOI: https://doi.org/10.1145/2531602.2531711

[37] Philipp Lorenz-Spreen, Stephan Lewandowsky, Cass R. Sunstein, and Ralph Hertwig. 2020. How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature Human Behaviour* 4, 11 (2020), 1102–1109. DOI: https://doi.org/10.1038/s41562-020-0889-7

[38] Gary Marchionini. 2006. Exploratory search: From finding to understanding. *Communications of the ACM* 49, 4 (2006), 41–46.

[39] Dana McKay, Stephann Makri, Marisela Gutierrez-Lopez, Andrew MacFarlane, Sondess Missaoui, Colin Porlezza, and Glenda Cooper. 2020. We are the change that we seek: Information interactions during a change of viewpoint. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 173–182.

[40] Paul Mena. 2020. Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook. *Policy & Internet* 12, 2 (2020), 165–183. DOI: https://doi.org/10.1002/poi3.214

[41] Boaz Miller and Isaac Record. 2013. Justified belief in the digital age: On the epistemic implications of secret internet technologies. *Episteme* 10, 2 (2013), 117–134. DOI: https://doi.org/10.1017/epi.2013.11

[42] Mohsen Mosleh, Gordon Pennycook, Antonio A. Arechar, and David G. Rand. 2021. Cognitive reflection correlates with behavior on Twitter. *Nature Communications* 12, 1 (2021), 921. DOI: https://doi.org/10.1038/s41467-020-20043-0

[43] Sean A. Munson, Stephanie Y. Lee, and Paul Resnick. 2013. Encouraging reading of diverse political viewpoints with a browser widget. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*.

[44] Raymond S. Nickerson. 1998. Confirmation Bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* 2, 2 (June 1998), 175–220. DOI: https://doi.org/10.1037/1089-2680.2.2.175

[45] Rik Peels. 2016. *Responsible Belief: A Theory in Ethics and Epistemology*. Oxford University Press.

[46] 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 (2019), 39–50. DOI: https://doi.org/10.1016/j.cognition.2018.06.011

[47] Richard E. Petty and John T. Cacioppo. 1986. The elaboration likelihood model of persuasion. In *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*, Richard E. Petty and John T. Cacioppo (Eds.). Springer, New York, NY, 1–24. DOI: https://doi.org/10.1007/978-1-4612-4964-1_1

[48] Gloria Phillips-Wren and Monica Adya. 2020. Decision making under stress: The role of information overload, time pressure, complexity, and uncertainty. *Journal of Decision Systems* 29, sup1 (Aug. 2020), 213–225. DOI: https://doi.org/10.1080/12460125.2020.1768680

[49] Tenelle Porter, Abdo Elnakouri, Ethan A. Meyers, Takuya Shibayama, Eranda Jayawickreme, and Igor Grossmann. 2022. Predictors and consequences of intellectual humility. *Nature Reviews Psychology* 1, 9 (2022), 524–536. DOI: https://doi.org/10.1038/s44159-022-00081-9

[50] Tenelle Porter and Karina Schumann. 2018. Intellectual humility and openness to the opposing view. *Self and Identity* 17, 2 (2018), 139–162. DOI: https://doi.org/10.1080/15298868.2017.1361861

[51] Suppanut Pothirattanachaikul, Takehiro Yamamoto, Yusuke Yamamoto, and Masatoshi Yoshikawa. 2020. Analyzing the effects of "people also ask" on search behaviors and beliefs. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*. ACM, New York, NY, 101–110. DOI: https://doi.org/10.1145/3372923.3404786

[52] Emmanuel M. Pothos, Stephan Lewandowsky, Irina Basieva, Albert Barque-Duran, Katy Tapper, and Andrei Khrennikov. 2021. Information overload for (bounded) rational agents. *Proceedings of the Royal Society B: Biological Sciences* 288, 1944 (Feb. 2021), 20202957. DOI: https://doi.org/10.1098/rspb.2020.2957

[53] Alisa Rieger, Frank Bredius, Nava Tintarev, and Maria Soledad Pera. 2023. Searching for the whole truth: Harnessing the power of intellectual humility to boost better search on debated topics. In *Proceedings of the Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–8. DOI: https://doi.org/10.1145/3544549.3585693

[54] 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*. ACM, 189–199. DOI: https://doi.org/10.1145/3465336.3475101

[55] Alisa Rieger, Qurat-Ul-Ain Shaheen, Carles Sierra, Mariet Theune, and Nava Tintarev. 2022. Towards healthy engagement with online debates: An investigation of debate summaries and personalized persuasive suggestions. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, New York, NY, 192–199. DOI: https://doi.org/10.1145/3511047.3537692

[56] Christina Schwind and Jürgen Buder. 2012. Reducing confirmation bias and evaluation bias: When are preference-inconsistent recommendations effective – and when not? *Computers in Human Behavior* 28, 6 (2012), 2280–2290. DOI: https://doi.org/10.1016/j.chb.2012.06.035

[57] Chirag Shah and Emily M. Bender. 2022. Situating search. In *Proceedings of the ACM SIGIR Conference on Human Information Interaction and Retrieval*. ACM, New York, NY, 221–232. DOI: https://doi.org/10.1145/3498366.3505816

[58] Boban Simonovic, Katia Vione, Dean Fido, Edward Stupple, James Martin, and Richard Clarke. 2022. The impact of attitudes, beliefs, and cognitive reflection on the development of critical thinking skills in online students. *Online Learning* 26, 2 (2022), 254–274. DOI: https://doi.org/10.24059/olj.v26i2.2725

[59] Annelien Smets, Lien Michiels, Toine Bogers, and Lennart Björneborn. 2022. Serendipity in recommender systems beyond the algorithm: A feature repository and experimental design. In *Proceedings of the 16th ACM Conference on Recommender Systems. CEUR Workshop Proceedings*. 44–66.

[60] Catherine L. Smith and Soo Young Rieh. 2019. Knowledge-context in search systems: Toward information-literate actions. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. ACM, New York, NY, 55–62. DOI: https://doi.org/10.1145/3295750.3298940

[61] Diana Tabatabai and Bruce M. Shore. 2005. How experts and novices search the web. *Library & Information Science Research* 27, 2 (2005), 222–248. DOI: https://doi.org/10.1016/j.lisr.2005.01.005

[62] Richard H. Thaler and Cass R. Sunstein. 2008. Nudge: Improving Decisions about Health, Wealth, and Happiness. Penguin, 14–38.

[63] Keela Thomson and Daniel Oppenheimer. 2016. Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making* 11, 1 (2016), 99–113. DOI: https://doi.org/10.1037/t49856-000

[64] 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, 3 (2005), 251–271. DOI: https://doi.org/10.1207/S1532785XMEP0703_2

[65] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185, 4157 (1974), 1124–1131. DOI: https://doi.org/10.1126/science.185.4157.1124

[66] Dáša Vedejová and Vladimíra Čavojová. 2022. Confirmation bias in information search, interpretation, and memory recall: Evidence from reasoning about four controversial topics. *Thinking & Reasoning* 28, 1 (2022), 1–28. DOI: https://doi.org/10.1080/13546783.2021.1891967

[67] Bas Verplanken, Pieter T. Hazenberg, and Grace R. Palenéwen. 1992. Need for cognition and external information search effort. *Journal of Research in Personality* 26, 2 (1992), 128–136. DOI: https://doi.org/10.1016/0092-6566(92)90049-A

[68] Ingmar Weber and Alejandro Jaimes. 2011. Who uses web search for what: And how. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*. ACM, New York, NY, 15–24. DOI: https://doi.org/10.1145/1935826.1935839

[69] Ryen White. 2013. Beliefs and biases in web search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, 3–12. DOI: https://doi.org/10.1145/2484028.2484053

[70] Ryen W. White and Resa A. Roth. 2009. *Exploratory Search: Beyond the Query-Response Paradigm*. Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool Publishers, 1–98. DOI: https://doi.org/10.2200/S00174ED1V01Y200901ICR003

[71] Wikipedia. 2021. Confirmation bias. Retrieved 01 May 2021 from https://en.wikipedia.org/wiki/Confirmation_bias

[72] Ben James Winer, Donald R. Brown, and Kenneth M. Michels. 1971. *Statistical Principles in Experimental Design*. Vol. 2. McGraw-Hill, New York, NY.

[73] Luyan Xu, Mengdie Zhuang, and Ujwal Gadiraju. 2021. How do user opinions influence their interaction with web search results?. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, New York, NY, 240–244. DOI: https://doi.org/10.1145/3450613.3456824

[74] Steven Zimmerman, Stefan M. Herzog, David Elsweiler, Jon Chamberlain, and Udo Kruschwitz. 2020. Towards a framework for harm prevention in web search. In *Proceedings of the 1st Workshop on Bridging the Gap between Information Science, Information Retrieval and Data Science (BIRDS 2020), Co-Located with 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Ingo Frommholz, Haiming Liu, and Massimo Melucci (Eds.), Vol. 2741, CEUR Workshop Proceedings, Xi'an, 30–46.

Received 10 May 2023; revised 18 September 2023; accepted 9 November 2023