Boosting Intellectual Humility to Mitigate Confirmation Bias during Web Search

Msc Thesis Computer Science & Engineering

Frank Bredius

TU Delft

Boosting Intellectual Humility to Mitigate Confirmation Bias during Web Search

Msc Thesis Computer Science & Engineering

Thesis report

by

Frank Bredius

to obtain the degree of Master of Science
at the Delft University of Technology
to be defended publicly on July 6 2023, at 14:00

Thesis committee:

Chair:	Geert-Jan Houben
Supervisor:	Sole Pera
Daily Supervisor:	Alisa Rieger
External examiner:	Cynthia Liem
Place:	Faculty of Electrical Engineering, Mathematics, Computer Science, Delft
Project Duration:	November 2022 - July 2023
Student number:	4575377

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

Faculty of EEMCS · Delft University of Technology

Abstract

Searching the web on debated topics, which are issues under active discussion and where individual opinions diverge, can be highly challenging. It requires users to approach their queries objectively, browse many resources, and accept a certain level of uncertainty, even if it conflicts with personal values. In particular, confirmation bias, the tendency to favor information that reinforces beliefs or attitudes, impacts how people gather and interpret information retrieved from search engines when searching on debated topics. It can have far-reaching effects, potentially leading to conflicts, extremism, and polarization.

In an innovative approach towards mitigating the negative effects of confirmation bias on web search on debated topics, we propose implementing a boosting intervention aimed at enhancing Intellectual Humility (IH) – an individual's ability to acknowledge the fallibility of one's own beliefs and the limits of one's knowledge while remaining open to learning from others' perspectives even if they differ from their own viewpoint. While previous research has highlighted the potential benefits of boosting IH as a means to mitigate confirmation bias, its impact on users' search behavior has yet to be explored.

Our work bridges this gap through two randomized preregistered user studies, gaining valuable insights into the effectiveness of IH-boosting interventions in mitigating confirmation bias. In the first study, we assessed the effect of three boosting interventions with different levels of complexity on users' context-dependent IH. In the second study, we examined the effects of these interventions on web search behavior.

The first experiment successfully demonstrated the effectiveness of the interventions in boosting participants' IH across all three treatment groups. However, applying these interventions to web search, no significant differences in search behavior were observed. Our exploratory findings reveal that both individual and environmental factors, including occupation, personal viewpoints, and search results order, shape the impact of IH-boosting interventions on online search behavior, with varying effects observed across different debated topics. We hope this study inspires, and is an initial basis for continued efforts to explore the multifaceted relationship between IH, information-seeking behavior, and responsible opinion formation, ultimately promoting a more informed and unbiased online discourse.

Preface

As I approach the end of my student journey at TU Delft, I reflect on the rewarding and educational experience that completing this thesis has been. From the very beginning, the weight of the final deadline loomed over me, creating a sense of both excitement and nervousness.

I owe a debt of gratitude to several individuals who have played a significant role in my thesis journey. Alisa and Sole have been exceptional mentors, generously sharing their time, expertise, and valuable insights. Through countless brainstorming sessions and engaging conversations, they guided me in determining the most effective research approach and helped me navigate the complexities of structuring my study. Their unwavering support and guidance were crucial in shaping the outcome of this dissertation.

I would also like to express my sincere appreciation to Cynthia for her dedication in reviewing my report. And I would like to thank my friends and family for their support throughout this journey. They have been a constant source of encouragement, feedback, and patient listening to my struggles and triumphs. Their understanding and patience were invaluable.

As I finish this chapter in my academic career, I am glad for the knowledge gained, skills polished, and personal growth I have experienced throughout this thesis. I continue on the next step of my journey with a mix of nostalgia and excitement, carrying with me the essential lessons and experiences gained during my time at TU Delft.

Contents

List of Figures

List of Tables

Part I

Preliminary Analysis

1

Introduction

In the current Internet age, individuals have access to vast amounts of information through web search. This information can be used to inform personal and professional decisions on debated topics, which are issues under active discussion and where individuals maintain differing viewpoints, such as vegetarianism and bottled water [1, 2, 3]. Individuals must inform and possibly form opinions amidst an information overload [4, 5]. Additionally, the spread of false information, asymmetric relationships between platforms and people, and low-quality and non-expert content make web search on debated topics particularly challenging [6]. These challenges, including individuals' limited time for web interactions, can make information seeking on the web for debated topics also cognitively demanding, increasing the likelihood of individuals exhibiting confirmation bias [7, 8, 9, 10].

Confirmation bias involves the inclination to seek out and give more weight to information that confirms one's preexisting beliefs or attitudes [11, 12]. Confirmation bias encompasses multiple aspects, such as ignoring disconfirming evidence, underweighting such evidence, and resistance to altering one's perspective [12]. These aspects can greatly influence the accuracy and fairness of decision-making processes [10], leading to various negative consequences. Some of the more notables include:

- Attitude polarization: subjects with different views strengthen their own attitude even though they are presented with the same data [11, 13].
- Belief perseverance: keeping the same belief even though disconfirming evidence has been given [1].
- Irrational primacy effect: giving more weight to earlier evidence [14].
- Illusory correlation: falsely believing that there is a relationship between two events [15].
- Informed decision impediment: hindering the user from making well-informed decisions [3, 16, 17]

Consequently, confirmation bias is likely to contribute to various problems that can harm human welfare, including inter- and intra-group conflict, ideological extremism, and polarization [10, 5, 18].

Given the potential negative consequences of confirmation bias in combination with web search on debated topics, there is a need to explore possibilities that may mitigate its effects. However, the practicality and feasibility of certain proposed interventions, such as gaming and nudging, may be limited. Gaming interventions, as suggested by researchers such as Sellier, Scopelliti, and Morewedge [19], Dunbar et al. [20], and Poos, Bosch, and Janssen [21], require a significant time investment. Nudging, which has been explored in various studies (e.g., Rieger et al. [22]), may be perceived as manipulative and may not always be effective, particularly over the long term or across different domains [23]. More importantly, nudging, by definition, modifies the choice architecture or the environment in which decisions are made rather than directly affecting the user [24]. As a result, once the nudge is removed, there is a tendency for old behaviors to reemerge [23]. This underlines the necessity for more sustainable and transparent alternatives.

A suitable alternative might be boosting, which empowers users by enhancing their decision-making capabilities, making it a more sustainable and transferable approach to mitigating confirmation bias [6, 23, 25]. Instead of manipulating the environment, boosting develops the user's skills, making them more resilient to bias even when transitioning across different contexts or domains [23]. Furthermore, the

[1] https://www.populismstudies.org/Vocabulary/confirmation-bias/

potential perception of manipulation associated with nudging is likely to be lessened as boosting focuses on skill development rather than environmental modification. One example of a digital boost is a simple self-reflection exercise to create resistance against online manipulation or the use of a decision tree as an instrument for determining the credibility of information found on the web [2]. Given the promising use of boosting, we propose a boosting intervention to improve users' ability to conduct unbiased searches on debated topics and achieve well-informedness. We propose a similar boosting intervention as one that was shown to be effective in improving people's ability to detect microtargeted ads [6].

The existing body of work, including the research proposed by Lorenz-Spreen et al. [6], as well as other similar interventions until now, has primarily emphasized empowering people to confront issues prevalent in online environments, such as misinformation and fake news [6, 26, 27]. These interventions are aimed at enhancing individuals' skills and abilities to deal with such external challenges in the digital environment. However, concurrently, the academic community has not devoted notable attention to interventions that focus on challenges that originate from within the users themselves, such as confirmation bias. Notably, such user-centric challenges, often internal in nature, can also greatly impact the quality of information-seeking behaviors and, by extension, the process of opinion-forming and decision-making [28, 29]. Through our work, we aspire to address these inherent flaws. By doing so, we hope to encourage more unbiased and well-informed search behavior, thereby fostering responsible opinion-forming and decision-making. This leads us to our research objective:

Research Objective

Extend the existing literature on boosting to focus on internal rather than external challenges.

One potential solution lies in boosting a trait known as Intellectual Humility (**IH**), which we define in the context of this research as *acknowledging one's limitation of knowledge and beliefs, recognizing that these beliefs may be biased, and being open to learning from others and considering alternative viewpoints* [30]. This encompasses the acceptance that personal beliefs might be flawed, along with an awareness of the limitations in the evidence supporting these beliefs and one's constraints in gathering and assessing relevant information [31]. Studies also suggest that IH holds the potential to counteract confirmation bias [32, 33, 34, 35]. With this understanding, IH emerges as a highly prospective quality to enhance to mitigate confirmation bias during web searches.

Unlike other boostable user-related factors, IH directly addresses the cognitive and emotional biases underpinning confirmation bias [33]. Individuals with higher levels of IH are more open to considering alternative viewpoints and less likely to cling to their existing beliefs in the face of new evidence [30]. Additionally, research indicates that IH can temper polarization and extremism, lower susceptibility to conspiracy theories, and enhance learning, discovery, and scientific credibility [34, 36]. By encouraging individuals to consider a more extensive range of perspectives and assess evidence in a more balanced, unbiased manner, IH could be pivotal in promoting more thoughtful and reflective web search behavior. We posit this could lead to improved decision-making and lower susceptibility to confirmation bias during web searches [37, 38, 39, 23, 25, 4]. However, in the current research landscape, the impact of boosting IH on web search behavior has yet to be explored. This leads us to the second research objective:

Research Objective

Investigate the relationship between Intellectual Humility and confirmation bias during Web Search Behavior.

Research Question 1

Does boosting participants' IH through priming, a trait IH questionnaire, or a combination of both interventions affect their level of IH?

[2] https://www.scienceofboosting.org/tag/digital/

To examine the effectiveness of boosting IH in reducing confirmation bias during web search, we conducted two randomized controlled pre-registered user studies. Study 1 aims to determine whether a short boosting intervention can effectively increase participants' degree of IH, reflected by **Research Question 1**. We boost the participants via three interventions. The IH prime is a prompt explaining the concept of IH and its advantages. The trait IH questionnaire is a tool designed to elicit thoughts about IH with the goal of promoting increased awareness of it. We assess this by measuring the participant's state IH, which is the level of IH in specific contexts, both prior to and following the different interventions.

Research Question 2

What is the impact of a boosting application on confirmation bias during web search?

Research Question 3

Can confirmation bias during Web Search be mitigated by boosting Intellectual Humility?

Based on the second research objective, in Study 2, participants are asked to complete a web search task under different conditions, with and without one of the boosting interventions. The results are compared to determine the effectiveness of each intervention in mitigating confirmation bias, with the goal of countering any potential negative consequences that follow from it. With Study 2, we aim to answer **Research Question 2** and **Research Question 3**.

The findings of this research project challenge the initial hypotheses and suggest that we did not find evidence for an effect on confirmation bias through the proposed boosting approach during web searches. While this outcome provides valuable insights into the complexity of addressing confirmation bias, it opens up new avenues for exploring alternative strategies to tackle various biases that exist. Furthermore, the study contributes to our understanding of IH and paves the way for future investigations into how IH can be influenced or modified to address biases in information processing. Despite the unexpected results, this research still holds practical implications for individuals conducting online information searches. By integrating appropriate interventions into search engines or other information-seeking tools, users may be guided toward a more balanced and diverse range of information. This is particularly relevant when individuals are seeking information to make important decisions in areas such as healthcare, politics, or financial planning.

The key contributions of this study are outlined as follows.

- Exploration, comparison, and empirical evidence of three proposed novel IH boosting interventions that were able to significantly increase state IH. The interventions aimed at boosting IH through a short prime and through an IH awareness-enhancing questionnaire.
- Exploration and empirical analysis of the relationship between different IH boosting interventions, confirmation bias mitigation, and web search behavior.
- Sharing of two pre-registrations of the user studies conducted, which ensure reproducibility of these studies (see Footnote 5 and 6).
- A rich dataset containing behavioral data from the two studies conducted. Firstly, it contains information on the effect of three different boosting interventions on state IH. Secondly, it contains data from search logs, measures of knowledge, attitude, mood, emotion, receptiveness to opposing views, and demographics, indicating the effect of the boosting interventions on confirmation bias and on web search behavior. Put together, this dataset contains information from more than 500 users, see OSF[3,4].
- Another contribution is that of a search interface software project. This tool can be conveniently modified for various use cases, particularly for experiments requiring the monitoring of search behavior.

[3] https://osf.io/xa4cn

[4] https://osf.io/cg7zs

The remainder of this manuscript is organized as follows. In Chapter 2, we review existing research on confirmation bias mitigation during web search, including the factors contributing to it and the interventions developed to reduce it. We provide a rationale for why the IH boosting intervention is being studied in the current research and why it is likely to be effective at mitigating confirmation bias. In Chapter 3 and Chapter 4, we describe the research design, sample, and procedures used to collect and analyze data. The chapters contain a description of the boosting intervention being tested, measures used to assess confirmation bias, and potential moderators. Note that the pre-registration of each study, which offers a concise overview of the elements involved in the separate experiments, can be found on OSF[5,6]. In Chapter 5, the ethics concerning both experiments are briefly discussed. The findings of the research, including the effect of the boosting intervention on confirmation bias and observed moderating effects, are presented with the statistical analyses in Chapter 6. We interpret the study's results in the context of the existing literature on confirmation bias and boosting interventions in Chapter 7. We summarize the research's main findings, draw conclusions, discuss limitations, and mention suggestions for future work in Chapter 8.

[5]https://osf.io/7nd4g

[6]https://osf.io/6mbrs

2

Literature Background

In this chapter, we discuss background literature and closely-related works that form the basis for the work discussed in this manuscript. We start by examining the relationship between web search and confirmation bias, explaining how the process of online information retrieval can inadvertently reinforce pre-existing beliefs and prejudices. This is followed by a review of the existing strategies for mitigating confirmation bias within the context of web search, assessing their effectiveness and limitations. We then explore the concept of Intellectual Humility and its potential as a countermeasure against confirmation bias. Throughout this review, we identify gaps in the current body of knowledge that motivate our research.

2.1. Web Search and Confirmation Bias

People often exhibit pronounced confidence in their beliefs, often beyond what is justified [31]. This phenomenon, known as confirmation bias, is a cognitive bias wherein individuals are inclined to seek out and prioritize evidence that aligns with their existing beliefs [11, 12]. It also involves the tendency to disregard or undervalue conflicting evidence and exhibit resistance to changing one's mind [12]. These aspects can significantly impact the fairness and accuracy of decision-making processes [10], potentially contributing to polarization and extremism [10, 5, 18].

The concept of confirmation bias becomes particularly relevant when individuals are navigating the internet on topics that are controversial or widely debated [1, 2]. Such topics, which we refer to as "debated topics," often involve complex issues with multiple perspectives, such as *Should people become vegetarian?* [40]. Amid the digital landscape's challenges–information overload, the spread of misinformation, asymmetric relationships between platforms and users, and the prevalence of low-quality or non-expert content [4, 6, 7]–confirmation bias can be particularly influential [28]. This complex environment can make web searches cognitively demanding, providing plentiful opportunities for confirmation bias to surface and impact search outcomes [3].

Confirmation bias can infiltrate multiple stages of the web search process, as represented by Kuhlthau's 'Information Search Process' model [41] (see Figure 2.1). For example, during the *Formulation* phase, individuals are likely to formulate search queries that align with their pre-existing beliefs, often framing queries positively [42, 43, 44, 45, 8]. Such a bias in formulating search queries can yield search results that are skewed towards confirming their beliefs, thereby potentially limiting exposure to different perspectives. For instance, we illustrate in Figure 2.2 how search results differ between two queries with different viewpoints on the same subject. Further along the search process, during the *Collection* phase, users may exhibit a tendency to spend more time on pages that align with their beliefs during the source consultation phase [46, 47, 8]. Moreover, the bias may even impact how individuals recall and interpret the information gathered [46, 47, 8]. The cumulative effect of these influences can compromise the accuracy, completeness, and objectivity of the information retrieved from the web search [48, 49].

In light of these potential influences, researchers have proposed certain strategies to enhance the accuracy and objectivity of information retrieval. These strategies–spending more time on searches, issuing more queries, browsing more webpages for comparison, checking the evidence supporting webpage content, and collecting more evidence–aim to counteract the effects of confirmation bias [50, 51, 52]. To put these strategies into practice effectively, users need to approach the search task in an "exploratory" manner, which focuses on knowledge acquisition and comparison [53, 3, 54]. During an exploratory search,

TABLE 2. Information search process (ISP).

Stages in ISP	Feelings Common to Each Stage	Thoughts Common to Each Stage	Actions Common to Each Stage	Appropriate Task According to Kuhlthau Model
1. Initiation	Uncertainty	General/ Vague	Seeking Background Information	Recognize
2. Selection	Optimism			Identify
3. Exploration	Confusion/ Frustration/ Doubt		Seeking Relevant Information	Investigate
4. Formulation	Clarity	Narrowed/ Clearer		Formulate
5. Collection	Sense of Direction/ Confidence	Increased Interest	Seeking Relevant or Focused Information	Gather
6. Presentation	Relief/ Satisfaction or Disappointment	Clearer or Focused		Complete

Figure 2.1: Kulthau's Information Search Process Model (Adapted from Kuhlthau [41]).

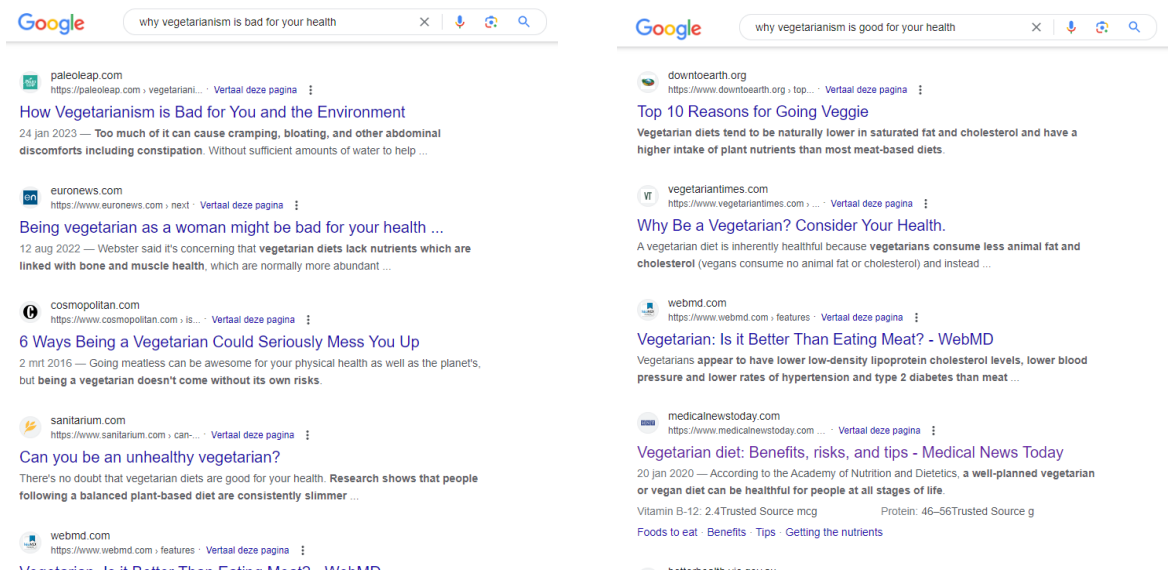
users are encouraged to gain a comprehensive overview of the topic by examining as many relevant items as possible [53]. This shift in search behavior could potentially mitigate the impact of confirmation bias, thereby fostering more balanced and objective information retrieval and promoting more informed decision-making.

2.2. Confirmation Bias Mitigation during Web Search

Confirmation bias mitigation has been the focus of researchers and practitioners in a broad range of domains, including psychology, cognitive science, and human-computer interaction [55, 20, 56, 57, 58, 59, 60]. Unfortunately, not all interventions are applicable to mitigate confirmation bias during web search. Consider for instance interventions that involve structured education programs and game-based environments, like training and gaming interventions respectively [19, 61, 20, 21, 62]. Both types of interventions provide the participants with strategies and skills that can help to reduce confirmation bias [19]. However, the time commitment required for these gaming and training interventions may render them impractical as a brief solution that mitigates confirmation bias during web searches, while maintaining a smooth web search experience.

On the other hand, nudging, a concept that refers to the practice of subtly guiding individuals' behavior through cues or prompts without imposing restrictions or mandatory rules [24], has gained significant attention in recent years. It is seen as a potentially effective method to influence behavior and mitigate confirmation bias during web search [22, 49, 56, 51]. The appeal of interventions that utilize nudging lies in their simplicity, and lack of time intensity, making them easy to deploy while being effective.

Several studies have investigated various nudging interventions aimed at mitigating confirmation bias during online searches. Rieger et al. [22] examined the impact of warning labels and task effects on interactions with attitude-confirming search results, although with potential harms to user autonomy. Draws et al. [63] explored whether exposure to viewpoint-balanced search results could shift attitudes, suggesting the possibility of a simple nudge within a web search context. Thornhill et al. [56] proposed two strategies–one to encourage users to explore further in their searches, and another offering quality feedback on shared posts–with promising outcomes for informed news consumption.



(a) Negative Viewpoint **(b)** Positive Viewpoint

Figure 2.2: Two results pages of a search engine, showing the difference in search results based on the viewpoint on a debated topic.

The field further extends with research by Schweiger, Oeberst, and Cress [49] into the reduction of confirmation bias through the provision of expert information and tag clouds with implicit evaluations. Similarly, Yamamoto and Yamamoto [51] proposed query priming to stimulate critical thinking and careful information-seeking through auto-completion and query suggestion. Other interventions, such as counter-argument techniques proposed by Huang et al. [64] and Huang, Hsu, and Ku [65], as well as alternative hypothesis consideration suggested by Koehler and Harvey [66], have also demonstrated effectiveness in reducing confirmation bias in seeking and evaluating new information. Collectively, these studies underscore the various approaches available for addressing confirmation bias, underlining the potential for effectively and easily promoting more balanced and critical information-seeking behaviors online.

While nudging has demonstrated the potential to reduce confirmation bias, it is important to recognize its limitations. The primary issue with nudging is that it focuses on altering the choice architecture or the decision-making environment rather than directly educating the decision-maker [24]. This can result in perceptions of manipulation and lack of transparency [23, 24, 67], possibly harming user autonomy [22]. Moreover, nudging may not always be effective, particularly over the long term or across different domains [23]. Therefore, these limitations underscore the need for alternative strategies that more directly and thoroughly tackle the individual's inherent cognitive biases.

Boosting is a policy approach that aims to improve individuals' decision-making skills and competencies rather than simply influencing their choices, as is done with nudging [23]. Leveraging principles from both behavioral and computational sciences, boosting interventions aim to enable individuals to make autonomous decisions [1]. Boosting applications have also resulted in a more lasting behavior change than nudging [23, 4]. And not less importantly, digital boosting interventions have successfully been applied during web interactions to help develop the necessary skills and knowledge to navigate the online world [2].

Boosting can be particularly valuable when governments do not always act in the best interest of citizens or where private sector companies create harmful choice architectures [23]. For example, search engines are unlikely to adjust their choice architectures if it would harm their commercial interests, making boosting a more desirable option for supporting generalizable and lasting behaviors [23]. In such cases, boosting may be a more effective approach to protect individuals and promote sound decision-making compared to nudging [23].

[1] https://www.scienceofboosting.org/about/

[2] https://www.scienceofboosting.org/digital

There are different types of boosting interventions, which we summarize in Table 2.1. To promote existing competencies or develop new ones, these interventions may employ human cognition, the environment, or both [3]. To qualify as a boosting intervention, specific criteria must be met [23, 4]:

- Interventions have to be quick, adaptive, and ethically acceptable.
- Interventions may build on current structures and utilize online world features.
- Interventions do not censor content but are based on context and knowledge of cognition.
- Interventions should aim to foster or develop new competencies under conditions of limited time and resources and typically in an adult citizenry that cannot be subjected to years of additional schooling.
- Interventions should preserve and enable individuals' agency and autonomy

Type	**Explanation**
Lateral leading	A straightforward guideline for verifying facts online [68]
Inoculation	Preventive intervention aimed at increasing individuals' resistance to online misinformation and misleading information. This is accomplished by exposing them to diluted disinformation and familiarizing them with typical deception strategies. **Games**: Bad News game [69], Cranky Uncle [70], Radicalise App [71]
Self-nudging	To improve self-discipline and reduce distractions by intentionally modifying an individual's immediate digital environment [25].
Critical ignoring	The deliberate process of screening and eliminating specific information in order to govern one's information sphere, decreasing the contact with incorrect or less beneficial material [72].
Simple Decision Aids	Simple strategies and tools aimed to improve people's ability to effectively evaluate the information they find online [73]

Table 2.1: Types of digital boosts targeting cognitive competencies.

In recent years, several intervention strategies have been proposed in existing research. Some researchers implement games as a boosting intervention [74, 26, 71]. These games are designed to enhance the ability of players to identify manipulation tactics. Examples include a browser game by Basol et al. [74] to recognize COVID-19 misinformation techniques and another by Roozenbeek and Linden [26] aimed at reducing the persuasiveness of fake news articles. The crux of these games lies in their different theoretical underpinnings. For instance, Saleh et al. [71] emphasize strategies derived from academic literature on extremism. While these games successfully improve participants' recognition of manipulative messages, they do not directly tackle cognitive biases like confirmation bias.

In addition to gaming, inoculation boosts, often through feedback provision, represent another intervention aimed at empowering users to make well-informed decisions [25, 23]. An example is Lorenz-Spreen et al. [6] research on closing the knowledge gap between users and microtargeting platforms. Their experiment showed that a simple boosting intervention could indeed enhance people's ability to detect targeted advertisements. Yet, like gaming interventions, these types of boosting primarily focus on raising awareness about external manipulative practices, leaving personal cognitive biases relatively unaddressed.

Counter-messages represent another boosting strategy, as outlined by Linden et al. [27]. They examined how consensus cues are processed in polarized information environments and found that preemptive warnings about politically motivated misinformation can protect public attitudes about scientific consensus. Still, similar to the interventions discussed previously, this approach focuses primarily on external misinformation rather than the user's cognitive biases.

Although significant progress has been made in boosting research in recent years, there remains a gap concerning user-centered problems such as confirmation bias. Current interventions primarily address external issues–misinformation and fake news [6, 26, 27]–and while they enhance users' abilities to navigate these challenges, they fall short in addressing internal cognitive biases. This highlights the need

[3] https://www.scienceofboosting.org/

for interventions that target cognitive biases directly, thereby fostering better decision-making skills. To fill these gaps, our focus turns toward developing brief boosting interventions that directly target cognitive biases such as confirmation bias, potentially making these interventions more widely accessible and effective.

2.3. Intellectual Humility

To address these cognitive shortcomings, we focus on the user's metacognitive state by addressing certain user-centric factors [20]. As shown in Table 2.2, several such factors have been identified in the literature as relevant for mitigating confirmation bias. Each factor comes with a brief description, accompanied by references to studies where this factor has been explored or utilized. This user-centric approach seeks to fill the gap identified in previous paragraphs, emphasizing interventions that directly target internal cognitive biases, thereby potentially enhancing their effectiveness.

User related factor	**Explanation**	**Source**
Confirmation proneness	To evaluate the tendency to exhibit confirmation bias.	[12, 20]
Cognitive flexibility	The ability to adapt our behavior and thinking in reaction to the environment	[75, 21]
Cognitive Reflection	The ability to overwrite an intuitive answer to a question with the less intuitive but correct answer	[76, 22, 21, 19]
Intellectual Humility	Openness, curiosity, tolerance of ambiguity, low dogmatism. IH Scale is a valid measure of the degree to which people recognize that their beliefs are fallible	[31, 33]
Need for closure	The degree to which a person has the desire for certainty. People who obtain high scores on this scale value order, dislike ambiguity, make decisions, form impressions quickly, and have strong opinions.	[20, 31, 77]
Need for Cognition	How much a person enjoys effortful cognitive activities.	[78, 20]
Big 5 personality traits	Openness, conscientiousness, extroversion, agreeableness, emotional stability	[20]
Actively Open-Minded Thinking	Measures the degree to which a person is willing to consider opposing viewpoints and change their mind about topics	[63]
Immersion	To measure the immersiveness that participants experience when playing a game or reading a story.	[21]

Table 2.2: A (non-exhaustive) list of factors researched together with confirmation bias or web search behavior.

Studies suggest that Intellectual Humility (IH) holds the potential to counteract confirmation bias and therefore emerges as a highly prospective quality to enhance in order to mitigate confirmation bias during web searches [32, 33, 34, 35]. Individuals with higher levels of IH are more open to considering alternative viewpoints and less likely to cling to their existing beliefs in the face of new evidence [30]. In addition, IH has been shown to reduce confirmation biases when reasoning about evidence and evaluating beliefs [33]. Moreover, IH has also been negatively correlated with political confirmation bias, indicating that individuals with higher IH are less likely to engage in biased reasoning when evaluating political topics [32, 37]. IH has been consistently linked with greater respect for and openness to opposing views [37]. On a societal level, IH may have important interpersonal implications for a person's perception and understanding of social extremism and polarization [79]. In addition, research suggests that IH can decrease polarization, extremism, and susceptibility to conspiracy beliefs, increase learning and discovery, and foster scientific credibility [34]. Overall, boosting IH is expected to contribute to making well-informed non-biased decisions [34].

Existing research has demonstrated that IH can be boosted through various interventions [37, 80, 34]. In the first study (see Chapter 3), we aim to develop an IH-boosting intervention tailored to the context of web searching. Furthermore, drawing on previous interventions, we seek to adapt and apply techniques

that have shown promise in enhancing IH. One approach to boosting IH is by changing people's mindsets. For example, studies have found that individuals who hold incremental theories of intelligence, believing that intelligence can grow and evolve, tend to exhibit higher levels of IH [37]. Experimental research has also demonstrated the applicability of this approach to state IH [80].

Various other strategies have also proven effective in boosting IH. Encouraging individuals to think from a distance perspective – that is, to step back and view a problem from a broader standpoint – has been found to increase levels of IH [80, 34]. Additionally, letting people explain a topic they perceive to possess significant knowledge can enhance IH. For example, studies have shown that individuals tend to overestimate their self-reported understanding of a policy less after writing a detailed explanation of how that policy works [34]. Reflecting on IH through questionnaires or engaging in brief reflection, writing, or reading exercises designed to shift IH has also been documented to yield short-term gains in IH [34]. Another effective approach to boosting IH is via priming. For instance, priming participants to consider the fallibility of their knowledge about a topic effectively enhances IH, particularly in individuals with high trait levels of self-reported state IH, though not as effective in those with low trait levels of state IH [81]. Simply reading about the benefits of intellectual humility, as opposed to high certainty, has also been found to boost self-reported IH [34, 38]. For instance, Porter et al. [38] conducted a study wherein the participant read an article emphasizing the advantages of acknowledging the limits of one's knowledge, see Figure 2.3 for a snippet of this article, or see OSF[4] for the full article. These studies suggest that IH can be temporarily boosted through simple, low-cost techniques, making them feasible for implementation in a web search context [34].

> It turns out that Bock's quite right in hiring this sort of employee. Dr. Harry Benzinger, head researcher at Harvard's Economic Institute, studied over 600 companies and found that those with a higher proportion of "knowers" did much better. "You might think that it's important to show what you don't know. But actually, **being very vocal in showing how much you know can have profound benefits. It makes others see you as an expert and leads to more rapid decision making, which ultimately promotes ingenuity,**" states Benzinger.

Figure 2.3: A snippet of the article that primes IH by highlighting the benefits, as used by Porter et al. [38].

A gap mentioned in the literature is the need for more exploration of how individuals become intellectually humble [37]. Although some interventions have been described, more interventions are needed to examine whether interventions to boost IH can address complex social problems such as polarization, misinformation, and conspiracy beliefs [34]. Furthermore, although IH has been identified as a potentially protective factor against confirmation bias and other forms of biased thinking, there is a need for more research to examine the relationship between IH and confirmation bias during web search specifically. An overview of the discussed literature and the gap we address in this research, as illustrated in Figure 2.4, serves to highlight the importance of our study and the specific contribution it makes to the existing knowledge in the field.

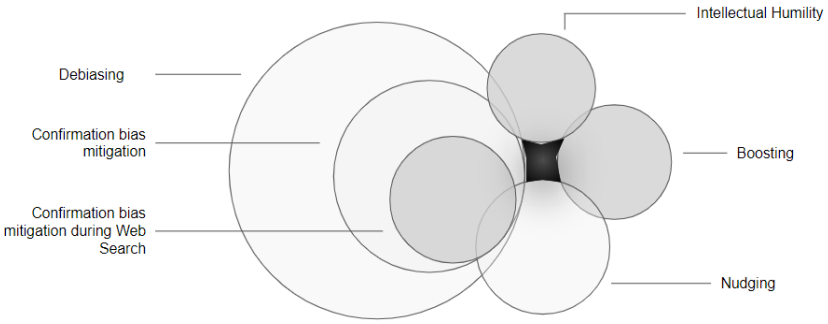


Figure 2.4: A Venn diagram showing the discussed literature and the research gap.

[4] https://osf.io/t5v3c

Part II

Implementation

3

Boosting Intellectual Humility

In this chapter, we describe the methods of the first study performed, which focuses on creating the proper boosting intervention that boosts IH. This chapter focuses on describing the study conducted to answer research question **RQ1**, which is *Does boosting participants' Intellectual Humility (IH) through priming, a trait IH questionnaire, or a combination of both affect their IH level?* The first intervention involves priming societal values related to intellectual humility to increase participants' awareness of its importance. The second intervention consists of a trait IH questionnaire to participants' awareness of their level of IH, intending to increase their level of IH. Finally, the third intervention combines both interventions, first the prime and then the questionnaire.

3.1. Design Plan

Study 1 followed a traditional experimental model, as described by Kelly [82], where baseline measures of outcome variables were taken before introducing the stimulus. The study utilized a between-subjects design with three groups, where treatment conditions were manipulated. Two hypotheses were formulated to answer **RQ1**.

H1: The state IH level of participants in the treatment groups will increase.

This hypothesis proposes a relationship between **Treatment Condition** (Independent Variable 1) and **State IH difference** (Dependent Variable 1), see Section 3.5. The interventions include an IH questionnaire and an IH prime. The IH prime is a prompt explaining the concept of IH and its advantages. The trait IH questionnaire is a tool designed to elicit thoughts about IH with the goal of promoting increased awareness of it (refer Section 3.2 for more explanation on the interventions). We also expected that participants' desire to present themselves in a favorable light, which can be influenced by social desirability bias [80, 83], may contribute to increased IH levels for participants doing the trait IH questionnaire.

H2: The full intervention group will have the highest increase in IH level, followed by the questionnaire group, and the prime group will have the lowest increase in IH level.

This hypothesis also proposes a relationship between **Independent Variable 1** and **Dependent Variable 1**. The full intervention, which includes both the IH prime and the trait IH questionnaire, allows participants the most significant opportunity to become aware of their behaviors and subsequently increase their State level of IH the most. Given that the Trait IH questionnaire is considerably more extensive than the brief IH Prime message, we expected that the questionnaire would result in a greater increase in State IH than the IH Prime.

3.2. Material

State Questionnaire We use the state IH questionnaire from Hoyle et al. [30] in our study as it has several advantages over other measures. This questionnaire allows us to assess IH levels in a specific context, capturing variations in individuals' intellectual humility as they navigate different situations [34]. Compared to trait measures, which may not be as sensitive to change, the state questionnaire is better at

detecting variability and capturing the impact of interventions [30, 34]. The items included in the state IH questionnaire can be found in Appendix A. The items in the scale contain blanks to be filled in with the specific issue of interest, i.e., the debated topic, see paragraph 3.2. The items were rated on a 1 to 7 scale, reflecting the degree of resonance between the participant and the given statements, with 1 denoting ”not at all like me” and 7 representing ”very much like me”. The resulting average score provides an index of the participant’s state IH.

Trait Questionnaire The goal of the questionnaire is to raise awareness about the participants’ values, which in turn should increase IH. The trait questionnaire is taken from Alfano et al. [84]. This questionnaire was chosen since it captures multiple dimensions of IH and is extensively validated. This scale is related to openness to new ideas, curiosity, low dogmatism, and tolerance of ambiguity. See Appendix A for the items displayed in the trait IH questionnaire. Each item was scored between 1 and 7, and then the mean of these scores was computed. Note that although the outcome of this questionnaire was not used as a dependent variable, we maintain an interest in the trait IH levels among participants in the Questionnaire and Full intervention treatment groups. To calculate trait IH, we scored the levels of agreement with the questionnaire statements from 1 (strongly disagree) to 7 (strongly agree). For negatively worded items, we applied reverse scoring. Finally, the average score for each of the four subscales was calculated to establish the overall degree of trait IH.

IH Prime The work of Lorenz-Spreen et al. [6] inspires the IH prime. A prime is a stimulus that activates a mental concept and influences subsequent behaviors [51]. This prime aims to increase IH by briefly describing what it is and its benefits. Research has shown that simply reading about the benefits of being intellectually humble boosted IH [37, 38]. See Appendix A for the text displayed as IH Prime.

Topics Before the IH questionnaires, the participants were given a list of topics for which they indicated their beliefs. One of these topics, for which the participant had a strong viewpoint, was then chosen to be filled in the blanks in the items of the state IH questionnaire (see paragraph 3.2). To guide topic selection, we relied on the following criteria, inspired by the works of Draws et al. [63] and Rieger et al. [22]:

- The topics need to be debatable, meaning that there are different opinions regarding the subject.
- Topics should have a somewhat equal distribution of opinions, e.g., climate change is not used as a topic given that there is clear scientific evidence.
- Topics should not be relevant only in a specific part of the world, e.g., gun legislation (USA).
- The topics need to be used in other literature on confirmation bias during Web Search.

The chosen topics are derived from the following sources. ProCon[1] presents sourced pros and cons of debatable issues. Moreover, Wikipedia[2] has a list of controversial issues. Chelaru et al. [85] also mention 50 controversial topics. And finally, Draws et al. [40], Rieger et al. [22], Alaofi et al. [86], Draws et al. [87], Potthast et al. [88], and Gezici et al. [89] share debated topics used in web search experiments on search behavior. In the end, the topics considered in this study are:

- Should people become vegetarian?;
- Is drinking milk healthy for humans?;
- Should students have to wear school uniforms?;
- Is homework beneficial?;
- Is obesity a disease?;
- Is Cell Phone Radiation safe?;
- Should bottled water be banned?;
- Should zoos exist?;
- Are social networking sites good for our society?

[1] https://www.procon.org/

[2] https://en.wikipedia.org/wiki/Wikipedia:List%20of%20controversial%20issues

3.3. Sample

Recruitment

We recruited participants from Prolific [3]. Each participant was allowed to participate in the study only once and should be older than 18, have an approval rate between 80 and 100 percent, and be fluent in English.

Sample Size

We anticipated observing medium effects of the boosting interventions on users' search behavior (cohen's $f = 0.25$). Assuming $f = 0.25$, a significance threshold α = $\frac{0.05}{2}$ = 0.025 (due to testing two hypotheses), a power of (1- β) = 0.8 and given that we tested up to three groups (i.e., three intervention conditions: *prime*, *questionnaire*, *full intervention*, we determined in a power analysis for a between-subjects one-way ANOVA with fixed effects (see Section 3.6) using the GPower software that the required sample size for this study is 189 participants.

3.4. Data

Procedure

The participants were randomly assigned to one of the three treatment groups: (1.1) Questionnaire group, (1.2) Prime group, and (1.3) Full intervention. Before the experiment, the participant was expected to enter their demographics. The collected demographics consist of their gender, age, educational background, and country of residence. See Table 3.1 for a brief overview of the differences between the treatment groups, and Figure 3.1 for the participants' workflow.

We asked the participants about their opinion regarding the selected debated topics (see paragraph 3.2). They were asked to state their agreement towards statements on a 7-point Likert scale ranging from *Strongly disagree* to *Strongly agree*. The topic for which the participant had the strongest opinion was selected. With this design choice, we tried to match conditions in which confirmation bias is most likely to occur [30]. The topic names were used in the IH state questionnaires throughout the rest of the survey.

1. In (1.1), the participant fills in the IH state questionnaire, the IH trait questionnaire, and then the IH state questionnaire again.
2. In (1.2), the participant fills in the IH state questionnaire, sees the IH prime, and then fills in the IH state questionnaire again.
3. In (1.3), the participant fills in the IH state questionnaire and is then exposed to both the IH trait questionnaire and the IH prime, after which the IH state questionnaire is shown.

Group	Pre-test	Condition	Post-Test
Questionnaire	Topics, IH state	IH trait questionnaire	IH state
Prime	Topics, IH state	IH prime	IH state
Reinforce	Topics, IH state	IH prime + IH trait questionnaire	IH state

Table 3.1: Overview of the different procedures for the different groups.

Collection

The data collection for this study took place over three days, from April 25^{th}, 2023, starting at 22:26, until April 28^{th}, 2023, concluding at 21:00. The participants, who were reimbursed at a rate of £1.05 for an estimated seven minutes of participation time, amounted to an overall cost of £264.60 for the experiment. This compensation rate equates to an hourly wage of £9.

The recruitment process was initiated with small participant batches of 10, allowing us to monitor the quality of the incoming data. Once we were confident in the integrity of the data and in determining the experiment's duration for accurate participant compensation, we increased the batch size to 40 after the first 30 participants. We ran the experiment at different intervals throughout the day to accommodate global participants in different time zones.

[3] https://www.prolific.co/

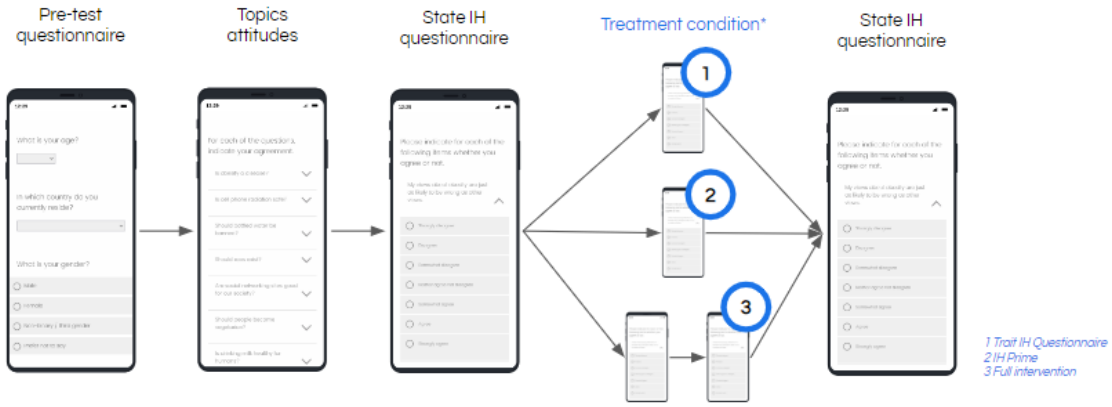


Figure 3.1: Study 1 participant workflow. Note that the study can be completed on different digital devices.

During the initial data collection stage, we discovered that the survey questionnaires lacked a validation requirement. This absence could have allowed participants to proceed without providing the necessary responses. Fortunately, this oversight did not impact the data, as all participants completed the questionnaires fully, and we rectified the issue promptly.

Dataset

The dataset generated from this study is composed of individual responses gathered via Qualtrics. The data contains various variables, including each participant's Prolific ID, their explicit consent, assigned treatment condition, demographic information (age, country of residence, gender, and education), Intellectual Humility (IH) measures (state IH before and after the intervention, state IH difference, trait IH), viewpoint strength on a scale of 0-3, their chosen topic viewpoint and corresponding ID, and an array of topic viewpoints. Additionally, the dataset contains information on whether participants passed attention checks and the total duration of their participation in the study.

Data Exclusion

Participants were excluded if they failed two or more of the (maximum) four attention checks. Attention checks were three for the Prime and Questionnaire groups and four for the Full Intervention group.

Preprocessing

The preprocessing of our dataset was executed utilizing the Python programming language. Initially, we performed a data-cleaning procedure that entailed the elimination of any incomplete or duplicate entries. This step ensures that only valid and unique data points contribute to our analysis. The desired data format was a CSV file containing columns for all relevant variables, with each participant represented by a single row.

3.5. Variables

Here, we outline independent, dependent, and exploratory variables used. Detailed information on how these variables were measured can be found in Section 3.2.

Independent Variables

IV1: Treatment Condition (between-subjects)

Participants were randomly assigned to one of the three conditions:

1. IH Trait Questionnaire
2. IH Prime
3. Full intervention (IH Prime + IH Trait Questionnaire)

Dependent Variables
DV1: State Intellectual Humility Difference (-3, 3)

Descriptive and Exploratory Measurements
The following variables were collected or calculated to provide descriptive and exploratory analysis.

1. Attitude measurement for each of the topics

   To capture participants' attitudes, we present them with a 7-point Likert scale on which they report their agreement to a statement on the topics, ranging from *Strongly disagree* to *Strongly agree*.
2. Trait Intellectual Humility

Demographics

1. Country of residence
2. Age
3. Gender

3.6. Hypothesis Testing

All analyses are performed in Python, and we uploaded the according code to OSF. Three t-tests for comparison between groups were applied. In the t-test, the IH State level difference (**H1, H2**) between (1) *IH Trait Questionnaire group*, (2) *IH Prime group* and (3) *Full intervention group* is compared. The significance threshold for hypothesis testing is set at alpha = 0.05 / 2 = 0.025 (Bonferroni-corrected due to testing two hypotheses). Outliers were not excluded from the analysis.

4

Boosting Intellectual Humility to Mitigate Confirmation Bias during Search

With Study 2, we aimed to investigate the impact of three interventions designed to boost IH on web search behavior. The three interventions consist of a prime, a trait questionnaire, and a combination of both, see Chapter 3 for a more detailed explanation. In the previous study, we found that these interventions effectively boost the state-level IH of participants. The aim of this study is to investigate whether the interventions have a positive effect on search behavior in terms of reduced confirmation bias and effortful search. By exploring the relationship between IH and web search behavior, this study aims to contribute to the understanding of this relationship and ultimately promote unbiased search practices that enable responsible opinion formation.

4.1. Design Plan

This study followed the classic interactive information retrieval design, wherein the baseline is introduced as an alternative to the experimental system, as described by Kelly [82]. It was conducted in the form of a randomized controlled trial between-subjects design, with four groups which where exposed to different **treatment conditions**. Also, additional factors were explored for new hypothesis-forming. This chapter focuses on describing the study conducted to answer **RQ2**, which is *What is the impact of a boosting application on search behavior?*, and on answering **RQ3**, which asks *Can confirmation bias during Web Search be mitigated by boosting Intellectual Humility?*. We formulated five hypotheses to answer these two research questions.

H3: Participants in the boosting condition will have lower *Click Proportion Attitude Confirming Results* compared to participants in the baseline.

Reasoning for H3: This hypothesis proposes a relationship between **Treatment Condition** (Independent Variable 1) and **Click Proportion Attitude Confirming Search Results** (Dependent Variable 1). Given that the interventions boost IH, and that prior research has shown that higher IH was consistently linked with greater respect for and openness to the opposing view [37], and increased engagement with attitude-opposing viewpoints [32, 90], we expect that the participants in the treatment conditions will show a decrease in the proportion of clicks on attitude confirming results, and thus perform a more thoughtful and balanced search behavior.

H4: Participants in the boosting condition will have greater *Lowest Rank Document Clicked* compared to participants in the baseline.

H5: Participants in the boosting condition will have longer *Average Dwelling Duration* compared to participants in the baseline.

H6: Participants in the boosting condition will have longer *Task Completion Time* compared to participants in the baseline.

H7: Participants in the boosting condition will have more *Cumulative Clicks* compared to participants in the baseline.

Reasoning for H4-7: These hypotheses propose a relationship between **Treatment Condition** (Independent Variable 1) and **Lowest Rank Document Clicked, Dwelling Duration, Task Completion Time**, and **Cumulative Clicks** respectively (Dependent Variables 2-5).

Existing literature suggests that browsing lower-ranked web search results can lead to the acquisition of more accurate information on the web [50]. In line with this, previous research has shown a positive correlation between higher levels of IH and increased motivation for knowledge acquisition, as well as more extensive engagement in information seeking and exploration of diverse viewpoints [37, 91, 92]. Building on these findings, we hypothesize that increasing participants' IH through the intervention will lead to more effortful searching, characterized by a propensity to explore lower-ranked search results, an increased number of cumulative clicks, longer dwell times on web pages and longer task completion times. By encouraging a more thorough and comprehensive search process, we anticipate that participants will have a greater likelihood of obtaining accurate information and forming responsible opinions.

H8: The *Full intervention* will have stronger effects on the user's search behavior than the *Questionnaire* and *Prime* interventions.

The *Full intervention*, which includes both the IH prime and the trait IH questionnaire, allows participants the greatest opportunity to become aware of their behaviors, and subsequently, we expected that it would have stronger effects on search behavior.

4.2. Data

4.2.1. Procedure

The participants were randomly assigned to the four treatment groups: (1.1) Questionnaire group, (1.2) Prime group, (1.3) Full intervention, and (1.4) Control. See Figure 4.1 for an overview of the workflow of the participants.

Pre-screening We asked participants to fill in a pre-test questionnaire. This questionnaire consists of items about their demographics and their viewpoint regarding a list of topics. The user was asked whether they strongly disagreed, strongly agreed, or something in between - on a 7-Point Likert scale.

Treatment condition Depending on the assigned condition, the participant was exposed to a different intervention before advancing to the search.

- In (1.1), the participant fills in the trait IH questionnaire.
- In (1.2), the participant sees the IH prime message.
- In (1.3), the participant is exposed to both the IH prime and the trait IH questionnaire.
- In (1.4), the participant won't get any intervention.

Search After the users had read the instructions (see **Search Task Instructions**) the users performed a search task on a mock search engine. They can enter a query, after which the search engine results page is shown. There they can click on the documents and view the contents. Their behavior using this search engine is logged. When they have finished searching, participants can click the *Stop searching* button, after which a completion code appears. This code should then be entered into the Qualtrics survey. See Appendix B for images showing the mock search engine.

Post-test Once the search task is completed, the participant is asked again to share their viewpoint on the topic, indicate their perceived knowledge gain on the topic, and can provide their arguments in a short text field. The primary purpose of this text box is to satisfy the participants, but it can serve exploratory purposes. Moreover, the participants are asked to answer some reflection questions regarding the experiment (see paragraph 4.4), and the NASA-TLX questions (see paragraph 4.4).

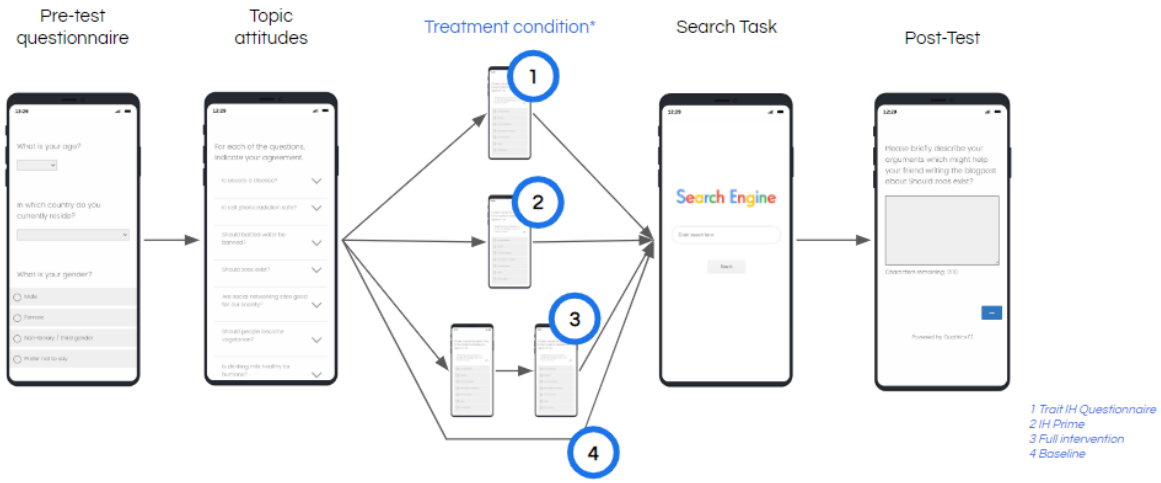


Figure 4.1: Study 2 participant workflow.

4.2.2. Collection

The process of data collection for this study was executed in multiple stages. The initial screening experiment started on May 8^{th}, 2023, at 10:09 and concluded the same day at 17:23. Participants were compensated at an average hourly rate of £6.28 for their involvement in the study, which required an average completion time of 1 minute and 26 seconds.

The main experiment was subdivided into six distinctive topics (Should people become vegetarian; Is drinking milk healthy for humans; Should students have to wear school uniforms; Is homework beneficial; Is obesity a disease; Should bottled water be banned), each constituting a separate experiment on Prolific. The initial experiment was launched on May 9, 2023, at 14:56, and the final experiment concluded on May 11, 2023, at 12:09.

The participant recruitment strategy began with small batches of 10 individuals for a single topic, enabling us to observe the quality and integrity of the preliminary data. Once satisfied with the data's quality, we expanded the batch size to secure a sufficient number of participants for each topic, approximately 70. To accommodate global participants, we staged the experiment at various intervals throughout the day, considering different time zones.

In retrospect, it was discovered that our data-logging process only recorded the use of the mouse button click on the search button on the search engine's homepage, failing to log instances where the ENTER key was used. This oversight resulted in a lack of data for 55 out of 352 cases concerning the exploratory variables of Query Length and First Query Duration.

4.2.3. Dataset

The dataset derived from this study contains individual participant data collected via Qualtrics, Prolific, and LogUI platforms. The data includes a variety of variables (please refer to Section 4.3 for an explanation of all the variables), such as participants' Prolific ID, explicit consent, the specific topic explored during the search, assigned treatment condition, demographic information (including language, age, country of residence, gender, and education), measures of Intellectual Humility (IH), metrics of interaction with different document types (clicks and time spent), values for each dependent and exploratory variable, the mean document rank, the participants' viewpoint on each topic, NASA-TLX values, participants' self-reflection on their search behavior, their comments on the study, their agreement with the topic after the experiment, and their knowledge gain on the topic post-experiment. Furthermore, the dataset includes data on whether participants successfully passed attention checks and the total duration of their participation in the study.

4.2.4. Data Exclusion
Following our preregistered exclusion criteria participants were excluded if they failed two or more of the four attention checks, or if they spent less than 40 seconds on the search engine. The attention checks were divided among the prime questionnaire, the trait IH questionnaire, and the topics pre-screening questionnaire, and an attention check was included in the search experiment.

4.2.5. Preprocessing
We processed our dataset using the Python programming language. Initially, the logs obtained from LogUI were downloaded as .log files and converted into a JSON format. Subsequently, we developed a script that extracted relevant information from these logs, such as clicks and hovers, and compiled them into a comprehensive dataset. This dataset was then merged with the Qualtrics datasets containing data from the pre-screening test and the search test. From this combined dataset, we calculated the values for each dependent variable and exploratory measure (please refer to Section 3.5 for detailed instructions on how to perform the calculations). The final result is a dataset with columns representing the dependent variables and exploratory measures, and rows corresponding to each participant.

4.3. Variables
Here we outline the independent, dependent, and exploratory variables used.

Independent Variables
IV1: Treatment Condition

Participants were randomly assigned to one of the fours conditions:

1. Questionnaire
2. Priming
3. Full intervention
4. Control

Dependent Variables
We have selected the following five dependent variables (DV) which are useful metrics of information search behavior, inspired by the works of Athukorala et al. [54], Rieger et al. [22], Pothirattanachaikul et al. [45], Pothirattanachaikul et al. [44], and Suzuki and Yamamoto [50]. Athukorala et al. [54] used **DV2-5** to explore whether participants conducted an exploratory type of search. These variables also enable us to categorize the search into biased and non-biased search, using **DV1** [22, 28], and enable us to distinguish effortful from effortless search (**DV2-5**).

- **DV1:** Click Proportion Attitude Confirming Search Results

  The choice of this dependent variable is supported by the literature, which suggests the existence of confirmation bias during web search. While there exists a variety of self-report measures to assess the susceptibility of showing biases [93, 94, 95], no reliable object measure has been reported so far [96]. On the other hand, confirmation bias can also be measured by monitoring and analyzing search behavior [45, 44]. Confirmation bias can be observed in an increased likelihood of interacting with search results that confirm pre-existing beliefs rather than alternative possibilities [22, 28]. Additionally, confirmation bias may manifest through the use of positive test strategies [28, 97], selective search assertion [8], and actively seeking information that confirms their hypotheses or beliefs [98]. Besides, research suggests that intellectually humble individuals tend to spend more time learning about attitude-opposing viewpoints [32, 99, 37].

- **DV2:** Lowest Rank Document Clicked

  This dependent variable measures the lowest document on the search engine results page (SERP) clicked on by the participant. Empirical research by Pothirattanachaikul et al. [52] has inspired the choice for this variable, which is further supported by the idea that responsible information seeking involves browsing through lower-ranked web search results [50].

- **DV3:** Dwelling Duration

  This dependent variable measures the average time users spent by users investigating clicked documents in seconds. It is inspired by the work of Athukorala et al. [54] and White and Horvitz [29], who identified various metrics to differentiate between exploratory and lookup search. Exploratory search corresponds to a search that results in more informed decisions, something that is correlated to a higher IH [34]. By measuring dwelling duration, we aim to capture the extent to which participants invest effort and time in their search process.
- **DV4:** Task Completion Time (in milliseconds).

  This dependent variable is inspired by the works of Athukorala et al. [54] and Yamamoto, Yamamoto, and Fujita [100], who identified completion time as one of the metrics to distinguish between exploratory and lookup search behavior. Furthermore, Suzuki and Yamamoto [50] argue that spending more time searching is necessary to obtain correct information on the web. In contrast to traditional IR research, which often values low completion time, we are interested in a longer completion time in this context. This is because a more effortful search, as indicated by a longer task completion time, may reflect a responsible search process.
- **DV5:** Cumulative Clicks

  This variable is also inspired by the works of Athukorala et al. [54] and Yamamoto, Yamamoto, and Fujita [100]. Additionally, Suzuki and Yamamoto [50] highlighted the importance of browsing multiple web pages for comparison to obtain correct information on the web.

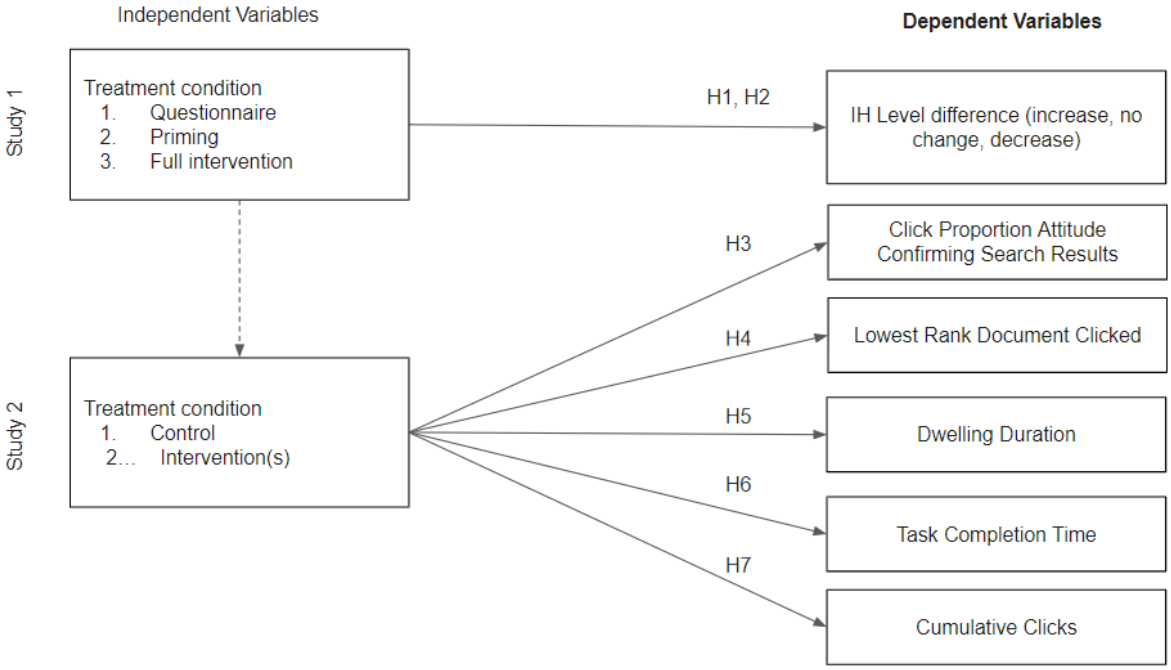


Figure 4.2: Overview of the (in)dependent variables for both studies.

Descriptive and Exploratory Measurements

In addition to the dependent variables, we have defined a list of descriptive and exploratory measurements aiming to more thoroughly understand the behavior and uncover new relationships and patterns. Collecting these varied measurements allows us not only to gain deeper insights but also to stimulate new research questions and hypotheses for future exploration. We have taken different specific, measurable factors from various literature about effortful and non-biased search, in order to be able to measure and study search behavior in a structured, systematic way [44, 52, 86, 97, 8, 54, 98]. Alongside these, we have

included topic measurements, which are necessary for the selection of an appropriate search task for the participant. The inspiration for these measurements originates from studies by Alaofi et al. [86] and Xu, Zhuang, and Gadiraju [98]. Furthermore, to ensure the diversity of our dataset, we have integrated demographic measurements.

Search Behavior

1. Number of queries issued in a search task.
2. Query Content (viewpoint and sentiment analysis)
3. First Query Length (amount of words).
4. First Query Iteration Duration. The time (in milliseconds) it takes between the first and second query, or equal to the total task time if only one query has been phrased.
5. Average time proportion spent on attitude-confirming documents. E.g. if the participants spend 30 seconds on attitude-confirming documents, and 30 seconds on attitude-opposing documents, the proportion is 0.5.
6. Maximum scroll depth in pixels related to the viewport of the browser.
7. Total Dwelling Duration in milliseconds.
8. Total Time Actively Spent on SERP in milliseconds.

Topics

1. Attitude measurement for each of the topics

   We present the participant with a 7-point Likert scale on which they can report their agreement to each of the topics, ranging from *Strongly disagree* to *Strongly agree*.
2. Self-reported knowledge about each of the topics

   We present the participant with a 7-point Likert scale on which they can self-report their knowledge on each of the topics, ranging from *No knowledge at all* to *Expert knowledge*.

Demographics

1. Country of residence
2. Age
3. Gender
4. Education

Search Consequences Post-search, participants filled out a questionnaire assessing their opinion changes, perceived knowledge gain, and new arguments formulated, collectively referred to as *search consequence* measurements.

1. Attitude change

   Attitude change will be calculated as the difference between pre-search and post-search attitude for the topic(s) a participant searched on. Positive values indicate a strengthening and negative values are a weakening of the initial attitude.
2. Knowledge gain

   Self-reported knowledge gain for the topic(s) a participant searched on.
3. Number of reported arguments

   The number of arguments a participant reported to have found during the search task.

Response categories Inspired by the works of Athukorala et al. [54] and Marchionini [53] we define three response categories used as exploratory measurements. Whereas a lookup search is described as a more basic kind of search with low objective complexity, exploratory search is described as an open-ended search with high objective complexity [54]. A search is categorized based on the values of dependent variables, where **DV1** indicates biased or unbiased search, and **DV2-5** indicates whether the search is either lookup or exploratory.

1. Lookup
2. Exploratory (unbiased)
3. Exploratory (biased)

4.4. Material

Below we discuss the material used in the experiment.

Search Task Instruction The phrasing of the task is inspired by [54, 47, 44, 22, 63]. Given that we wanted the participant to get new insights regarding a debatable topic, the search task is an exploratory search task with an open-ended search goal [53, 54]. Athukorala et al. [54] suggest that users behave differently depending on the search goal. Therefore, the task difficulty is fixated at a moderate level [54]. By phrasing the task in a neutral way we prevent it from influencing the experiment so that the search task type can be correctly measured. Note that in the instructions, *TOPIC* is replaced with the topic for which the participant had indicated to have a strong opinion. The search task instruction reads:

> A friend is telling you about a discussion they had with a colleague about *TOPIC*. The conversation made you curious. To learn more about *TOPIC*, you have decided to conduct a web search.
>
> Read the following steps carefully:
> 1. Click on the URL to open the custom search engine.
> 2. Explore the search results to learn more about *TOPIC*. You do not have to study every single search result, just use this search result page as you would normally.
> 3. Once you think that you have explored the search results enough, click the STOP button on the search engine.
> 4. Paste the code that you will receive below.
> 5. After completing the search, we will ask you to report the arguments you have found during the search task.

Topics and Search Results For the search task, we required viewpoint-annotated search results for multiple debated topics. For this, we use the dataset by Draws et al. [87], which contains search results for the nine topics described in paragraph 3.2. These viewpoint annotations of the documents are necessary to measure *Click Proportion Attitude Confirming Search Results* (**DV1**).

With these nine topics, we ran a screening with 650 participants on Prolific to identify eligible participants. We found 514 participants with strong opinions on three or more topics. We then optimized the topic selection to ensure that we have a somewhat equal distribution of participants among the included topics (i.e., we excluded topics for which only a few participants reported having a strong opinion).

This led to the following final topic selection:

- Should people become vegetarian?
- Is drinking milk healthy for humans?
- Should students have to wear school uniforms?
- Is homework beneficial?
- Is obesity a disease?
- Should bottled water be banned?

Search Results Page During the search sessions, we displayed ten documents per search result page for the assigned topics after a query had been entered and the search button had been pressed. To make the search process more realistic, a text similarity API[1] is used to compare the query with the search topic on a semantic level, which is optimized on short sentences. If the two sentences are not at all related (i.e. *Text Similarity* is smaller than 0.1 on a scale of 0 to 1), then we showed the message ”Could not find search results”. When the participant clicked on a search result, the web page opened in a new tab. Participants were expected to come back to the search engine after reading the document to continue the search task.

During the search task, the participants were exposed to three different versions of alternating order of viewpoints for the documents in the SERP, named *opposing*, *neutral*, and *confirming*. By doing this, we could measure and confirm that ranking bias/position bias did not influence the outcome of the experiment. To illustrate, when a user has filled in an agreement of *strongly disagree* towards the following statement: ”Bottled water should be banned.”, the user is put into one of the three groups. In this case, if the user is put in the *opposing* group, the participant will eventually see the *pro-SERP* (since the opposing opinion of disagree is pro). See Table 4.1 for the three different SERPs concerning the document viewpoint order.

document rank	**con**	**neutral**	**pro**
1	con	neutral	pro
2	neutral	pro	con
3	pro	con	neutral
4	con	neutral	pro
5	neutral	pro	con
6	pro	con	neutral
7	con	neutral	pro
8	neutral	pro	con
9	pro	con	neutral
10	con	neutral	pro

Table 4.1: The three types of SERPs with different viewpoints regarding each document.

Interventions The three pre-search IH-boosting interventions *prime*, *questionnaire*, and *full* are the same as those described in Chapter 3.

NASA Task Load Index The NASA-TLX test, a self-assessment tool used to estimate overall perceived workload [101], assists us in determining if individuals across varying conditions (like treatment conditions or age groups) perceive tasks distinctively, which may reveal interesting patterns. It captures a person’s workload over multiple dimensions, including mental demand (very low - very high), physical demand (very low - very high), temporal demand (very low - very high), performance (perfect - failure), effort (very low - very high), and frustration level (very low - very high). We excluded the item on physical effort and ask participants to respond to the following statements on a scale from 0 to 100 with 5-point steps:

1. How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex?
2. How much time pressure did you feel due to the pace at which the tasks or task elements occurred? Was the pace slow or rapid?
3. How successful were you in performing the task? How satisfied were you with your performance?
4. How hard did you have to work (mentally and physically) to accomplish your level of performance?
5. How irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the task?

From the participant’s responses, we calculate an unweighted average.

[1] https://dandelion.eu/docs/api/datatxt/sim/v1/

Attitude and Knowledge Gain To evaluate if our intervention effectively prompts participants to reassess and possibly alter their established opinions, we tracked changes in their attitudes. Meanwhile, self-reported increases in knowledge could signal a better search process, possibly due to a less biased approach leading to broader exposure to diverse information. For information on our attitude measurement methodology, refer to section Section 3.4. Participants were asked to rate their perceived knowledge accumulation during the search session on a five-point Likert scale, ranging from 'no knowledge gain' to 'substantial knowledge gain'.

Search Task Environment
To implement the search task environment, we created a website that resembled popular search engines, such as Google, but that was fully under our control. This allowed us to manage the documents returned by the search engine and monitor participant behavior. The architecture of the system contains several components, as illustrated in Figure 4.4. The system included various interfaces, accessible to either the administrator or the participant, as depicted in Figure 4.3.

LogUI LogUI is a framework used for logging participant behavior on the search engine. It consists of both a server application and a client implementation. In the client, it is specified which HTML elements and events are logged while the server application maintains a websocket connection with the client. The logs are stored on the server application, which runs in a Docker container. Logs captured by the LogUI server are saved in an SQLite database and can be downloaded in JSON format from the admin panel, ready for parsing by the Python data preprocessing code.

Django A URL, containing query parameters for the variables *prolific id*, *topic id*, and *viewpoint group*, is generated within the Qualtrics questionnaire. We developed a mock-up search engine using Django, featuring an index page with a query field and a search button. The SERP displays ten documents with their title, subtitle, and URL, along with navigation buttons to move between pages. Django was deployed using Gunicorn, and the search engine was built using HTML, CSS, JavaScript, and Python. The search engine also checked text similarity between queries and topics, providing a more realistic experience. This was accomplished via an HTTP call to the Dandelion API[2].

Server Both the LogUI application and the Django application were hosted on a virtual Linux server provided by TU Delft. We used Nginx to create three reverse proxies for the applications and the websocket, enhancing security by keeping internal deployment ports hidden from external access. Certbot was employed to create SSL certificates, further increasing the experiment's security.

Qualtrics Qualtrics was used to create the survey participants were required to complete. By incorporating JavaScript in Qualtrics, we employed logic to determine the appropriate SERP group, indicating the order of viewpoints shown on the SERP. For example, a participant supporting a certain topic in the confirming SERP group would see a SERP where the first document also supports the topic.

4.5. Sample

4.5.1. Recruitment

We recruited participants from Prolific[3]. Each participant was allowed to participate in the study (including all search sessions) only once, should be at least 18 years old, have an approval rate between 80 and 100 percent, and be fluent in English.

4.5.2. Sample size

We anticipated observing medium effects of the boosting interventions on users' search behavior (cohen's $f = 0.25$). Assuming $f = 0.25$, a significance threshold α = $\frac{0.05}{6}$ = 0.0083 (due to testing six hypotheses per search session), a power of (1- β) = 0.8 and given that we were testing, depending on the hypothesis, up to 4 groups (i.e., five intervention conditions: *control*, *prime*, *remind*, *reinforce*, we determined in a power analysis for ANOVAs (see Section 3.6) using the GPower software that the required sample size for this

[2]https://dandelion.eu/

[3]Prolific: `https://www.prolific.co/`

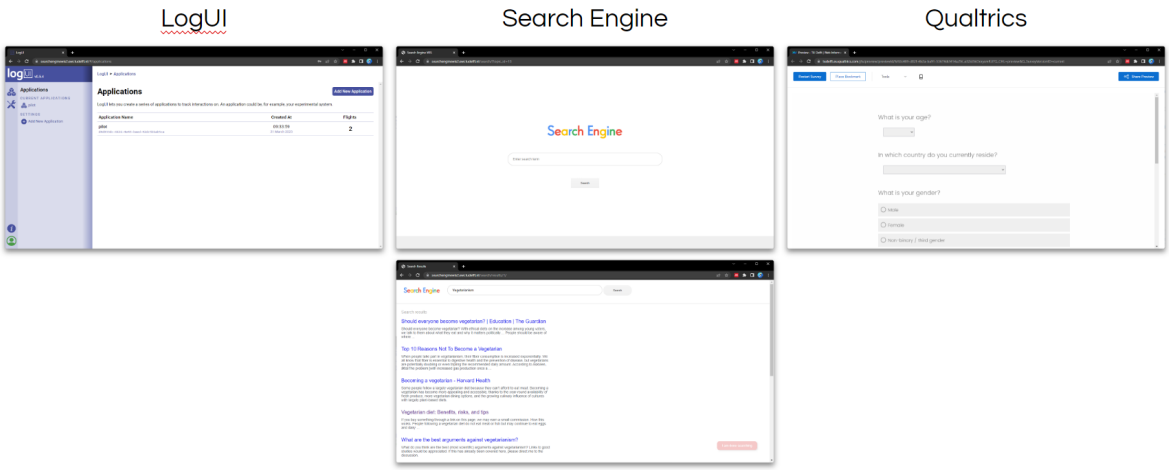


Figure 4.3: Different views of the system.

study is 285 participants.

4.6. Hypothesis Testing

All analyses were performed in Python, and we uploaded the according code to OSF. A one-way ANOVA and a one-way MANOVA for mean comparison between groups are applied. In the one-way ANOVA the difference in *Click Proportion Attitude Confirming Search Results* (**H1**) between (1) *Questionnaire group*, (2) *Prime group*, (3) *Full intervention group*, and (4) *Control group* is compared. In the one-way MANOVA the difference in web search behavior (*Lowest Rank Document clicked*, *Average Dwelling Duration*, *Task Completion Time*, *Cumulative Clicks*, **H2-5**) between (1) *Questionnaire group*, (2) *Prime group*, (3) *Full intervention group*, and (4) *Control group* is compared. The significance threshold for hypothesis testing is set at alpha = 0.05 / 5 = 0.01 (Bonferroni-corrected due to testing five hypotheses).

4.7. Exploratory Analysis

With the exploratory analysis, we aim to offer a deeper understanding of why our interventions led to the observed effects. The observed results could be attributed to a variety of factors, including noise in the data, the (in)effectiveness of our interventions in influencing behavior, or aspects of the experiment setup, such as the design of the search task. By examining a range of exploratory variables, we aim to provide insights into each of the factors and how the variables may influence participants' approach to search.

Workload: NASA-TLX Building upon the work of Kim [102], who emphasized the positive relationship between pre-task difficulty and web search interactions such as page viewing [103], we sought to find any noteworthy connections in our study. We delved into the impacts of the different interventions on participants' perceived workload. By assessing the effects of these interventions on workload, we aimed to gain insight into their potential role in reducing cognitive load and enhancing search efficiency. Furthermore, we explored the relationship between trait IH and perceived workload.

Demographics As stated by Weber and Jaimes [2], user demographics are among the most important predictors of online information search behavior, guiding us to consider the individual characteristics that may contribute to variations in search behavior. Moreover, based on the study conducted by Zou et al. [103], which revealed variations in search behavior based on education, we direct our attention to the student status of our participants.

Human Centered Factors We also focused on aspects that are intrinsically tied to individual user characteristics, such as trait IH, self-reported knowledge gain, the strength of the participant's viewpoints, and query sentiments. By exploring trait IH, we aimed to uncover insights into how inherent tendencies

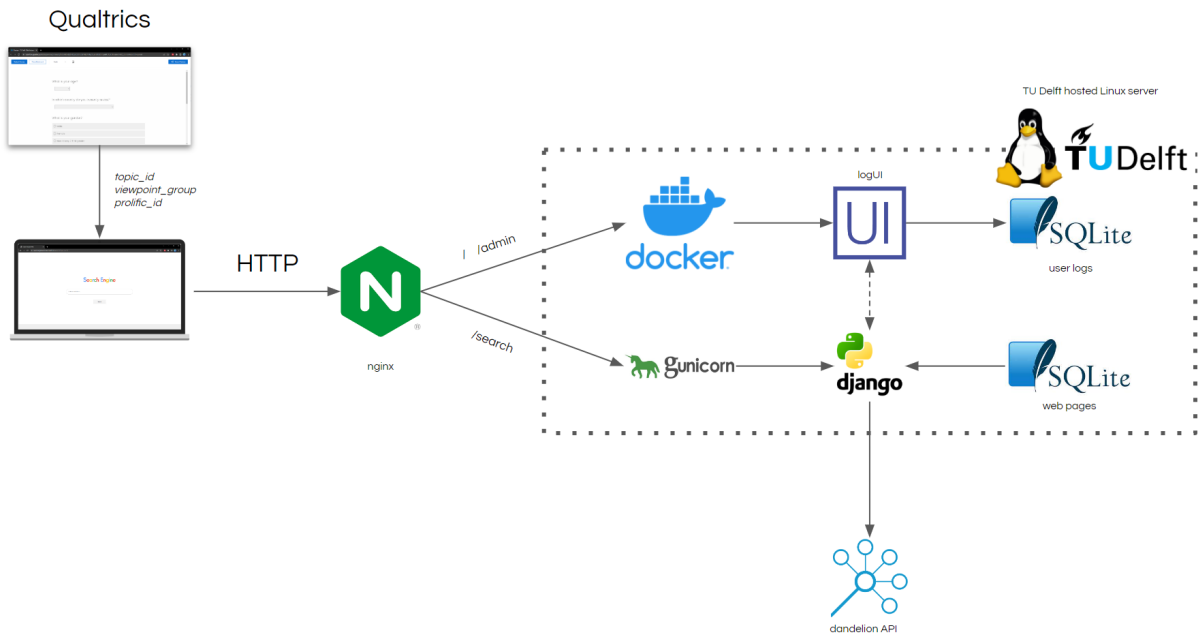


Figure 4.4: Architecture overview of the search engine.

toward openness and humility may impact search behavior in our experiment, as Porter and Schumann [37] also highlighted a negative correlation between trait IH and confirmation bias.

Task Related Factors Finally, we examined aspects that were related to the design of the experiment, such as the topic on which the participants had to perform the search task. Drawing from the proposition of Pothirattanachaikul et al. [52] that search engines can influence users' polarization by presenting belief-consistent results at higher ranks, we also analyzed participants' search behaviors across different SERP groups and treatment conditions.

5

Ethical Considerations

The implementation of studies involving human participants requires careful attention to ethical considerations. These considerations were examined and approved by the Human Research Ethics Committee before the beginning of the study.

5.1. Data Management

Our data management plan, approved by the data steward, outlined the specific strategies for handling our data. The plan addressed the types of data employed in the study (quantitative and qualitative questionnaire data, demographic information), the expected data volume (less than 250 GB), the intended storage locations both during the project (Project Storage at TU Delft) and after its conclusion (OSF), and the specifics of personal data collected (gender, age, consent form, country of residence). We also identified the category of our data subjects (crowd-workers) and the institutional affiliations (TU Delft). Our plan specifies that upon the conclusion of the research, all data was anonymized.

5.2. Ethics

Our ethical considerations followed a comprehensive checklist for human research, including a risk assessment and mitigation strategy. To minimize the risk of data breaches, we anonymized the Prolific IDs of participants. We ensured transparency by informing participants about their compensation before their participation, following the UK minimum wage standards as per Prolific's guidelines. Recognizing the potentially sensitive nature of some collected data (age, gender, country of residence, personal viewpoints), we required explicit consent from participants through a consent form. We obtained approval from participants to publicly release the dataset, excluding any personally identifiable information.

Our consent form clarified our study's purpose, procedures, potential risks, and voluntary nature. We assured participants that their responses would be kept confidential to the best of our ability, with data stored securely in password-protected electronic formats. We also assured participants that any data eventually published would be anonymized to prevent identification. This study received the approval of TU Delft's ethics committee, further substantiating our commitment to upholding ethical standards throughout our research.

6

Results

In this chapter, we present the findings of the research, including any statistical analyses that were conducted. It includes a description of the effect of the boosting interventions on both IH and web search behavior, as well as any moderating effects that were observed.

6.1. Study 1, Boosting IH

In this section, we discuss the results, findings, and analyses found in the data gathered through Study 1.

6.1.1. Descriptive Statistics

The study began with 190 participants, out of which 185 successfully passed the attention checks, establishing a viable sample for further analysis.

The demographic composition of the participants is as follows:

- Age distribution: The majority (81) of the participants were between the ages of 25-34, followed by 54 participants aged 18-24, 37 participants aged 35-44, 12 participants aged 45-54, and 6 participants aged 55-64.
- Country of residence: The participants were primarily from the United Kingdom (59), followed by Poland (38), Portugal (15), South Africa (12), the United States (12), Italy (10), Hungary (7), Spain (7), Germany (6), and Greece (6), with smaller representations from other countries.
- Gender: The majority of participants identified as male (123), while the remaining identified as female (67).
- Education: A variety of educational backgrounds were represented with 71 participants holding a 4-year degree, 41 being high school graduates, 40 having some college education, 26 possessing a professional degree, 7 with a 2-year degree, 3 with a doctorate, and 2 having less than high school education.

The participants were fairly evenly distributed across the three treatment groups, with 65 in the Full intervention group, 63 in the Questionnaire group, and 62 in the Prime group.

Participants were assigned to one of nine topics based on their strongest expressed opinions. The distribution across topics was as follows: obesity is a disease (71), cell phone radiation is safe (37), bottled water should be banned (22), zoos should exist (19), milk is good for your health (11), students should have to wear school uniforms (9), homework is beneficial (6), social networking sites are good for society (6). Notably, 58% of participants had a very strong opinion (either strongly agree or strongly disagree) about at least one of the statements.

Participant feedback on the experiment was generally positive, with comments like ”very straightforward,” ”Enjoyable,” and ”Thank you. I enjoyed the concise nature of this survey and its ability to capture my immediate thought.” Some participants also acknowledged the awareness-raising aspect of the experiment, with one stating, ”I did realize that I am lacking knowledge on the subject and that I was a bit biased about it.”

6.1.2. Hypothesis Testing

The first analysis, to test **H1**, involved the application of one-sample t-tests for each group to confirm the potential impact of the treatment conditions on state IH. The mean state IH (before) score was 4.197 (median: 4.222, SD: 1.181), indicating a relatively balanced distribution of intellectual humility prior to the experiment. The mean state IH (after) score was 4.510 (median: 4.667, SD: 1.261), suggesting a slight increase in IH after the experiment. The null hypothesis stated that the mean difference in state IH was zero, implying the treatments had no effect. We anticipated observing medium effects of the boosting interventions on users' state IH (cohen's $f = 0.25$). Given that two hypotheses were being tested, the significance threshold was adjusted accordingly, to alpha = 0.025, using Bonferroni correction. The t-value was calculated for each group using the formula:

$$t = (mean\_difference - 0)/standard\_error$$

Subsequently, the corresponding p-value was determined with $n - 1$ degrees of freedom. For the Full Intervention, Prime, and Questionnaire groups, the calculated t-values were 5.24, 6.04, and 2.77 respectively, see Table 6.1. Given these high t-values, the resulting p-values were small, all below the alpha level of 0.025. Hence, the null hypothesis was rejected in each case, indicating that all three interventions had a significant effect on state IH. These findings verified Hypothesis 1, demonstrating an increase in the state IH level of participants across all three groups.

Condition	Mean	Standard error	t	df	p-value
Full	0.3760684	0.0717	5.2433	64	1.894e-06
Questionnaire	0.223545	0.0560	2.7761	61	0.007265
Prime	0.3387097	0.0805	6.0479	62	9.71e-08

Table 6.1: T-Test values for each Treatment Condition.

To verify if there are statistically significant differences in the state IH between the three interventions (**H2**), a one-way ANOVA test was conducted. See Figure 6.1 for a visualization of the difference in state IH before and after the intervention per treatment group.

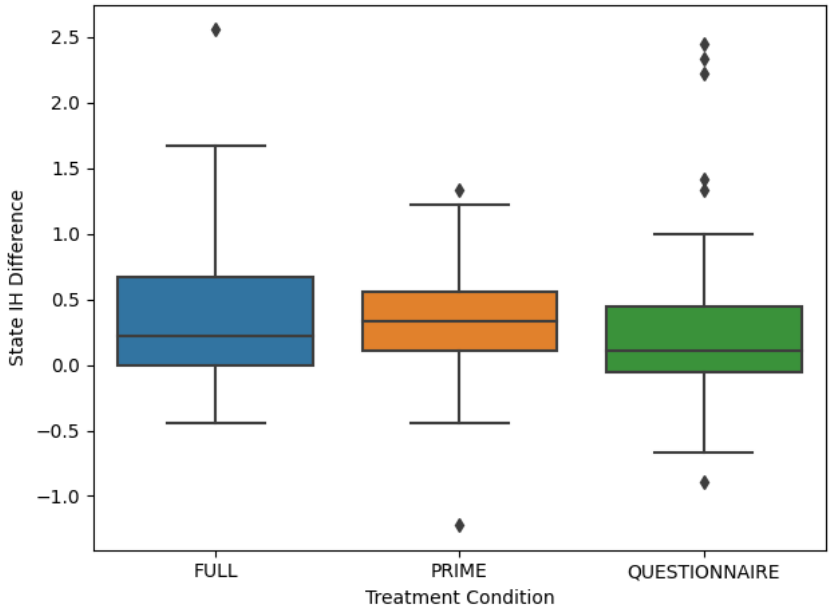


Figure 6.1: The difference in state IH before and after the intervention, per treatment group.

The null hypothesis (H0) for the ANOVA test is that the means of the state IH difference for the three groups are all equal. The alternative hypothesis (H1) is that at least one group's mean is different. After applying the one-way ANOVA test to the three groups, the F-value was not significantly high, and the

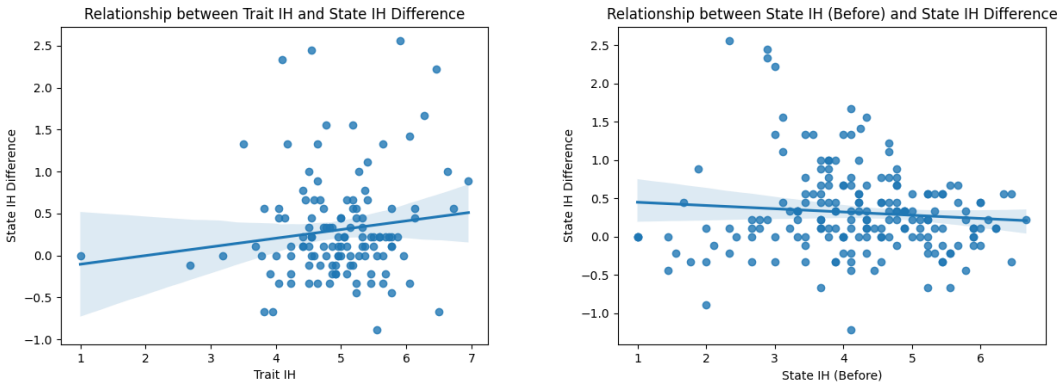
p-value was greater than 0.025 ($F = 1.282$, $p = 0.280$). Therefore, the null hypothesis could not be rejected. This suggests no significant difference in the state IH difference between the three groups.

6.1.3. Exploratory Findings

To uncover patterns and relationships that could provide valuable insights into the impact of the interventions, we delved deeper into the dataset. Our goal was to explore additional factors beyond the initial hypothesis tests and gain a deeper understanding of the effectiveness of the interventions. We focused on two key factors: participant characteristics and topics. While we observed some interesting trends, it is important to note that further investigation is required to draw definitive conclusions due to the limitations of sample size and the absence of specific hypotheses.

Participant Characteristics

Our initial step involved analyzing a range of participant attributes to ascertain their influence on the success of the interventions. Firstly, we explored the relationship between participants' initial state IH scores and the state IH difference. However, no significant correlation was found between the two variables. Similarly, the analysis did not demonstrate a clear relationship between trait IH scores and the state IH difference (see Figure 6.2). These findings suggest that pre-existing levels of IH do not appear to affect one's responsiveness to an IH-boosting intervention. In other words, individuals with both high and low initial IH levels are equally likely to be influenced by the intervention. This implies a broad potential for the intervention's impact across diverse IH levels.



(a) Trait IH

(b) State IH (Before the intervention)

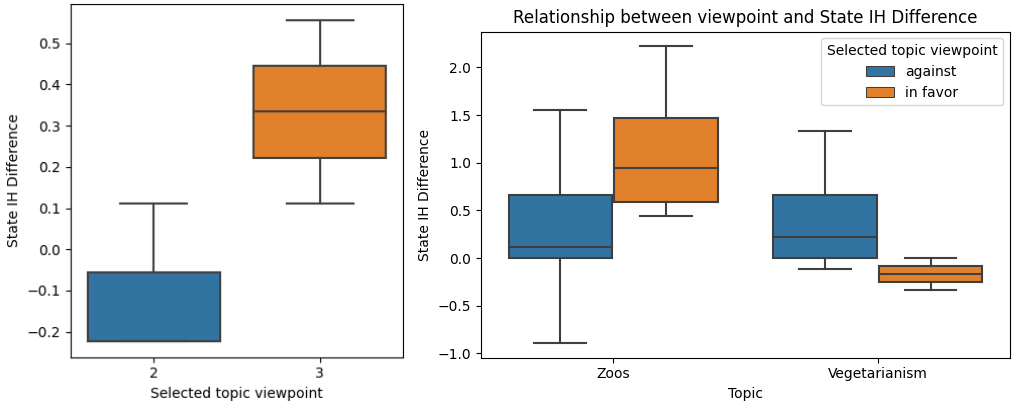
Figure 6.2: Relationship between state IH difference and IH.

In light of the absence of a significant relationship between state IH and trait IH with the effectiveness of the interventions, we further examined whether the strength of participants' opinions played a significant role (see Figure 6.3a).

Upon analyzing the data, we made an interesting observation in the context of the Social Network group. Participants with a strong initial opinion experienced an increase in state IH following the interventions, suggesting a positive impact. Conversely, participants with a less strong opinion exhibited a decrease in state IH (t-test, $t = -2.6186$, $p = 0.0601$). This suggests that individuals with strong convictions could be more open to IH-enhancing interventions, potentially because these interventions provide an opportunity to challenge and reassess strongly held beliefs. The apparent correlation between the strength of opinion and change in state IH also highlights the potential for more personalized, targeted interventions.

Considering these differences in viewpoint strength, we proceeded to investigate the role of participants' viewpoints in relation to the interventions. Specifically, we were interested in showing whether participants with opposing views would experience distinct impacts from the interventions. Upon analyzing the data for each topic, we observed some differences in the Zoos and Vegetarianism groups. Participants in favor of the statement appeared to experience the interventions differently compared to those who held opposing views (see Figure 6.3b) (2 t-tests, $t = 2.31, -2.44$, $p = 0.046, 0.07$). This indicates that the same intervention might have different impacts based on an individual's personal viewpoints. Potential factors

contributing to these differences could range from emotional ties to the topic, existing knowledge, or even cognitive dissonance triggered by conflicting information.

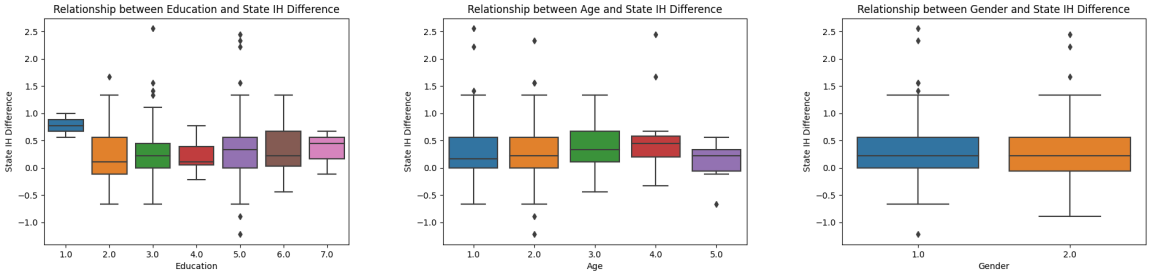


(a) Viewpoint Strength for topic Social Networks

(b) Viewpoint

Figure 6.3: Relationship between state IH difference and Participant Viewpoint.

Furthermore, we explored demographic factors, including age, country, education, and gender, to gain insight into their potential influence on state IH difference (see Figure 6.4). Differences based on the demographic factors between the participants were observed only for the demographic factor of education when grouping them. While these differences did not reach statistical significance, we did notice a variation in the relationship between education and the state IH difference. Specifically, participants with an educational background classified as ”Less than High School” (Education = 1) exhibited a greater increase in state IH difference compared to participants with other educational backgrounds. This suggests that education level might influence the susceptibility to an IH-boosting intervention.



(a) Education

(b) Age

(c) Gender

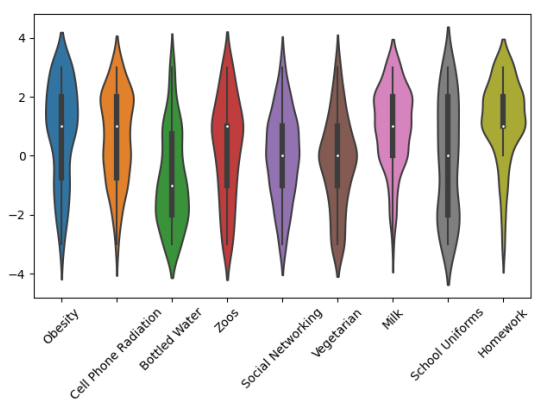
Figure 6.4: Relationship between state IH difference and Demographics.

Topic Characteristics

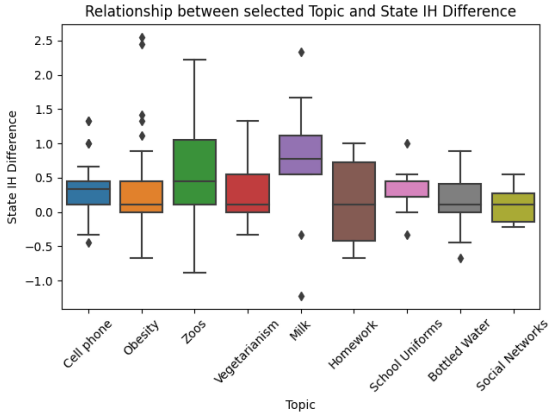
With some notable findings emerging from our exploration of participant characteristics, we shifted our attention to the topics. We explored how different topics influenced the effectiveness of the interventions. This analysis allowed us to find interesting patterns and relationships that provided insights into the varying impact of the interventions across topics.

Firstly, we examined the average viewpoint strength across topics, which ranged from 0 to 3 (*mean* = 0.997, *SD* = 0.539). There was a significant variation in viewpoint strength across the different topics, as shown in Figure 6.5 (a). Our observations indicate that there might be a difference between the topics Milk and Zoos and the other topics, see Figure 6.5b (e.g., t-test between topic Milk and Social Networks, $T = 1.5798, p = 0.1454$). This suggests that the nature of the topic could influence the effectiveness of interventions, and emphasizes the importance of considering the topic specificity and the user's viewpoint strength when designing and evaluating interventions aimed at boosting IH.

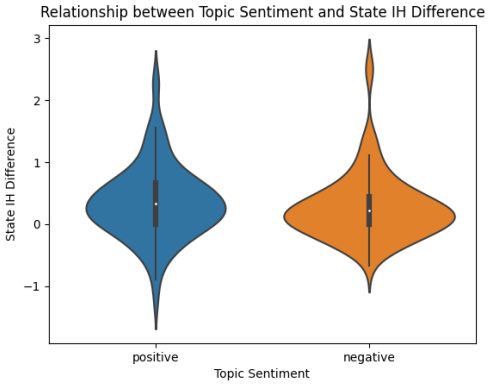
We also explored whether the sentiment of the topics (positive or negative) affected the interventions' effectiveness. For this, we used the lexical resource provided by Baccianella, Esuli, Sebastiani, et al. [104]. Using this resource, we categorized the topics based on their sentiment. To illustrate the categorization based on sentiment, we can consider the classification of the topics "Is cell phone radiation safe?" as a positive statement and "Is obesity a disease?" as a negative statement. However, we did not observe a considerable difference in the state IH difference between positive and negative sentiment (see Figure 6.5c).



(a) Topic Agreement



(b) State IH difference and selected Topic



(c) State IH difference and selected Topic sentiment

Figure 6.5: Exploration of Topic characteristics.

6.2. Study 2, Mitigating Confirmation Bias

In this section, we discuss the results, findings, and analyses found in the data gathered through Study 2.

6.2.1. Descriptive Statistics

Informed by prior longitudinal studies (see for example [105]), we anticipated a corrected planned sample size of 130% (370 participants) as we planned to conduct a second and third search session if we observed effects of decreased confirmation bias and/or increased search effort in the first search sessions. From the pre-screening sample (see paragraph 4.2.1), 432 participants reported to have strong opinions on three or more of these six topics and were thus considered to be eligible for the search study. We observed the participation of a diverse group of 352 individuals. However, to maintain the integrity of our study, we subjected the collected data to attention checks, which resulted in the exclusion of 11 participants. Thus, the analysis was conducted based on 341 valid entries.

The demographic composition of the participants is as follows:

- Age distribution: The age of the participants varied widely, from under 20 years to over 60 years,

with a significant majority between 21 and 30 years (188).

- Country of residence: The participants resided in various countries around the globe, with South Africa (77) and the United Kingdom (76) being the most represented, followed by Poland (51), Portugal (37), Italy (23), Mexico (11), and Spain (10), with smaller representations from other countries.
- Gender: The majority of participants identified as male (192), while the remaining identified as female (157).
- Occupation: Most of the participants held full-time jobs (171), and 116 participants were students.

Our participants were assigned to four treatment conditions: Control, Prime, Questionnaire, and Full intervention. These groups were well balanced in terms of participant count, with each ranging from 67 to 72 individuals. In addition, we established a Dummy Control group of 67 participants. This group was designed to serve a dual purpose. Firstly, it was meant to counter any potential argument that the interventions worked solely due to the longer duration of the experiment compared to the control group. To ensure this, the Dummy Control group was provided with an ATI questionnaire [106], similar to the intervention groups. Secondly, it was established to add an extra layer of robustness to our results by providing an additional comparison baseline. However, since the behavior of participants in the Control group was found to be identical to the Dummy Control group, we opted not to include this group in the final analysis. We reasoned that its inclusion would only add unnecessary complexity to the results without contributing significant new insights.

Participants were assigned to one of the six topics if they reported having a strong viewpoint on them. The distribution across topics was as follows: should bottled water be banned (64), is obesity a disease (61), should people become vegetarian (59), should students have to wear school uniforms (57), is homework beneficial (57), is drinking milk healthy for humans (53).

The distribution of participants was also equally balanced among different viewpoint ranking templates. These groups are determined by the highest-ranked result that appears on the SERP with a detailed explanation available in Section 4.4 concerning the different SERPs. In the *Opposing* group, we had 119 participants who encountered a result contradicting their beliefs as the first result. Meanwhile, the *Confirming* group consisted of 117 participants, and the *Neutral* group included 116 participants.

Participant feedback was also collected at the end of the experiment. The comments varied, ranging from a majority of positive feedback to observations on the sensitivity of certain topics. This feedback served as a qualitative meter of the experiment's reception among participants.

6.2.2. Hypothesis Testing

In the Hypothesis Testing phase of the second experiment, a comprehensive statistical analysis was undertaken to compare the different intervention groups, namely the Questionnaire group, the Prime group, and the Full intervention group, with the Control group. All statistical analyses were executed in Python, and the code has been made accessible on OSF for transparency and replicability.

The analysis included both a one-way ANOVA and a one-way MANOVA. The one-way ANOVA was employed to compare the *Click Proportion Attitude Confirming Search Results* between the intervention and the baseline groups, to test the first hypothesis of Study 2 (**H3**). The one-way MANOVA was used to compare web search behaviors, consisting of *Lowest Rank Document Clicked*, *Average Dwelling Duration*, *Task Completion Time*, and *Cumulative Clicks*, between the five different groups. These comparisons aimed at testing the remaining four hypotheses (**H4-7**). In the interest of adjusting for multiple comparisons and mitigating the risk of Type I errors, a Bonferroni correction was applied to the significance threshold, setting it at $alpha = 0.01$.

The results of the first analysis indicate that we did not find a statistically significant effect of the *Treatment Condition* on *Click Proportion Attitude Confirming Results*, which is shown by the F statistic of 0.3776 and an associated p-value of 0.82462. The F statistic and the p-value indicate that the boosting interventions did not have a significant effect on reducing participants' tendency to click on attitude-confirming search results. Thus, H3 was not supported. This lack of statistically significant effect is illustrated in Figure 6.6. This violin-graph visualization of our data reveals that the distribution of clicks within the intervention and control groups show no notable difference, confirming the statistical analysis.

When testing the remaining hypotheses with the MANOVA, the overall model's test statistics revealed a highly significant effect ($F = 34, p < 0.0001$), indicating that the outcome variables significantly differ from

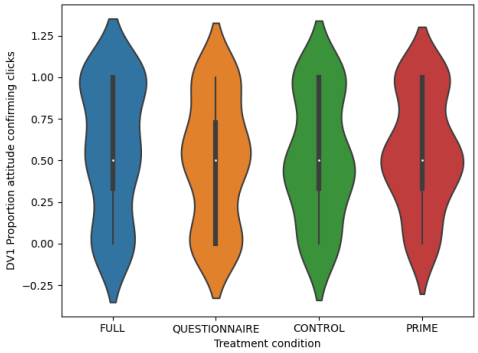


Figure 6.6: Visualization of the values for the dependent variable *Click Proportion Attitude Confirming Results.*

zero when the treatment condition would be at its baseline level, see Table 6.2. However, when evaluating the impact of the treatment condition on each of the dependent variables, the p-values associated with the F statistics were all above the adjusted significance threshold of 0.01, see Table 6.3. This suggests that we did not find evidence for an effect of the boosting interventions on the *Lowest Rank Document Clicked, Average Dwelling Duration, Task Completion Time, and Cumulative Clicks*. Therefore, our remaining hypotheses (**H4-7**) were also unsupported. The boxplot visualization of the data, illustrated in Figure 6.7, provides a clear picture of this data, aligning well with the results from our tests.

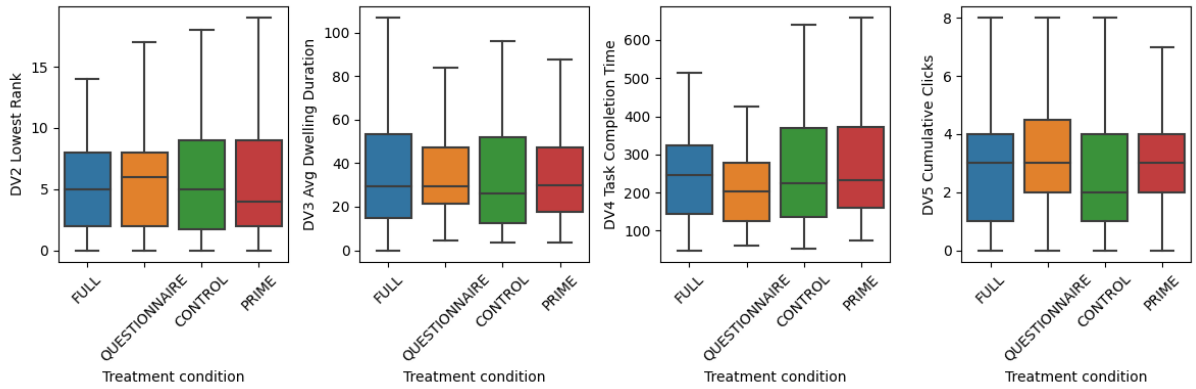


Figure 6.7: Boxplots of the relations between the treatment conditions and the dependent variables Lowest Rank, Average Dwelling Duration, Task Completion Time, and Cumulative Clicks.

Intercept	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.6885	4.00	302.00	34.1646	0.00
Pillai's trace	0.3115	4.00	302.00	34.1646	0.00
Hotelling-Lawley trace	0.4525	4.00	302.00	34.1646	0.00
Roy's greatest root	0.4525	4.00	302.00	34.1646	0.00

Table 6.2: MANOVA intercept table for hypothesis testing **H3-7**.

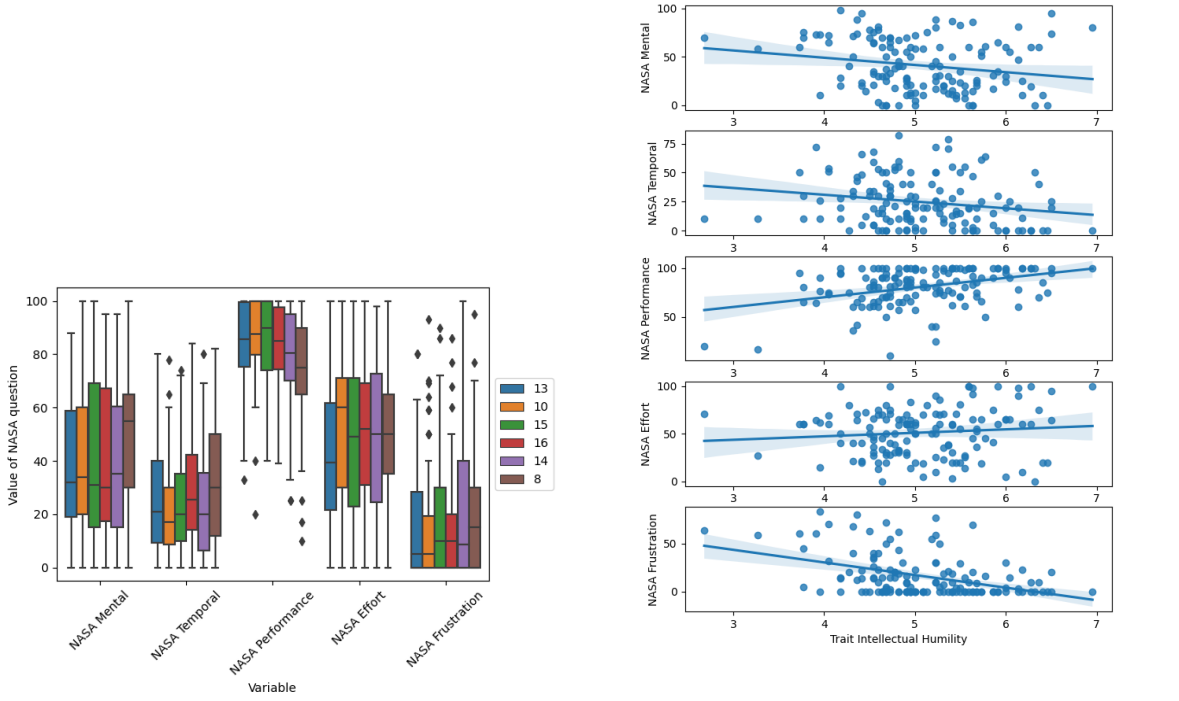
6.2.3. Exploratory Findings

Through our exploratory analysis, the objective is to get a more profound comprehension of the factors leading to the observed effects of our interventions. The following section will present our findings concerning each of the investigated factors, as outlined in Section 4.7.

Treatment_condition	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.9615	16.0000	923.2629	0.7468	0.7466
Pillai's trace	0.0389	16.0000	1220.0000	0.7487	0.7447
Hotelling-Lawley trace	0.0397	16.0000	598.0265	0.7465	0.7465
Roy's greatest root	0.0256	4.0000	305.0000	1.9495	0.1022

Table 6.3: MANOVA Evaluation of treatment condition on search effort.

Workload: NASA-TLX Our analysis revealed no significant trends linking perceived workload and search effort. Furthermore, we examined the individual NASA questions for each treatment condition. The results of the analysis shown in Figure 6.8a suggest that the topic "Is obesity a disease?" (Topic 8) influenced participants' perceived workload differently compared to other topics (One-way ANOVA results for each NASA question, $F = 6.619, 5.787, 13.295, 0.107, 0.761$, $p = 0.011, 0.016, 0, 0.744, 0.384$). When observing the relationship between trait IH and perceived workload, we found that participants with higher trait IH scores tended to report a lower workload (see Figure 6.8b). Specifically, participants with higher trait IH scores reported lower NASA Mental, NASA Temporal, and NASA Frustration scores, indicating fewer difficulties in cognitive processing. These participants also demonstrated higher NASA Performance and NASA Effort scores.



(a) Visualization of NASA-TLX subscales for each topic

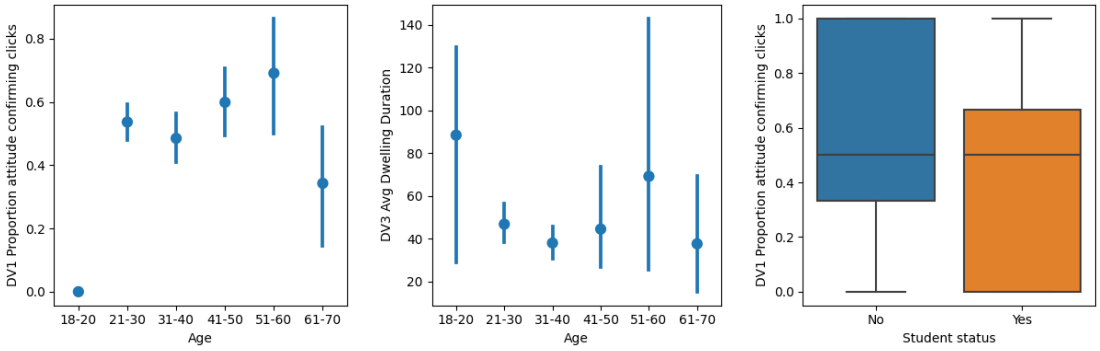
(b) Relationship between trait IH and self-reported workload

Figure 6.8: Exploration of perceived workload.

Demographic Factors Our analysis of the data revealed interesting insights regarding the influence of age on participants' search behavior. We observed that individuals in the age groups of 10-20 and 60-70 exhibited a decreased inclination to click on attitude-confirming results compared to other age categories (see Figure 6.9a). Additionally, participants in the 18-20 age group demonstrated a notably longer average dwelling duration (18-20 $mean = 88.44sec$, $std = 52.78$, other participants $mean = 45.23sec$, $std = 60.24$)

during their search tasks (see Figure 6.9b).

We observed that student participants displayed a decreased inclination to click on attitude-confirming results and demonstrated a more thorough and diligent approach to their searches (see Figure 6.9c). These observations prompted us to conduct a more detailed analysis of the treatment conditions to ascertain whether the interventions exacerbated these differences. Upon closer examination, we discovered that student participants in the intervention groups, particularly those in the Prime group, exhibited higher cumulative clicks compared to other student groups. However, it is important to note that across all student groups, there was a reduced proportion of clicks on attitude-confirming results, indicating a limited impact of the interventions in mitigating confirmation bias among students. Remarkably, our investigation further revealed that students assigned to the topic of School Uniforms exhibited the lowest proportion of clicks on attitude-confirming results.



(a) Relation between age and confirmation bias

(b) Relation between age and dwelling time

(c) Relation between education and search behavior per topic

Figure 6.9: Relationship of demographics with search behavior.

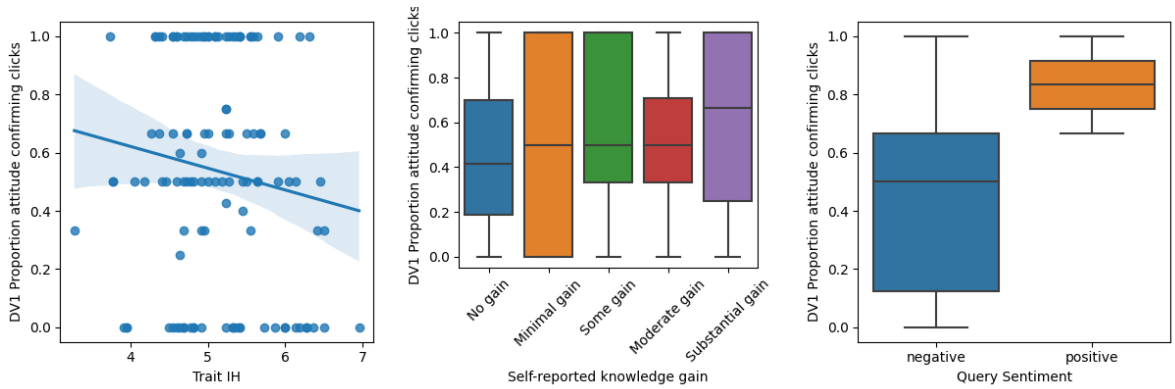
Human Centered Factors

Our analysis revealed that participants with a higher trait IH demonstrated a slightly lower proportion of attitude-confirming clicks (see Figure 6.10a). However, our exploration did not identify a discernible correlation between the level of trait IH and search effort.

An examination of the relationship between participants' self-reported knowledge gain and search behavior revealed interesting patterns. Participants who indicated no knowledge gain demonstrated the lowest proportion of clicks on attitude-confirming results ($mean = 0.46$, $std = 0.39$), see Figure 6.10b. Conversely, participants who indicated some knowledge gain exhibited increased search effort, indicating a more engaged and thorough information-seeking process. Notably, participants indicating no knowledge gain demonstrated the least search effort, potentially reflecting a lack of motivation or interest in exploring the topic further. We found no significant differences in self-reported knowledge gain across different treatment conditions.

The analysis of search behavior based on participants' viewpoint intensities and treatment conditions revealed no significant variations in search behavior. Participants with differing viewpoints on the task-related topic did not exhibit notable differences in their search behavior. Similarly, participants in different treatment conditions, each with varying viewpoint strengths on the topic relevant to their search task, did not display discernible differences in their search behavior.

The analysis of participants' query sentiment revealed no significant differences in the proportion of attitude-confirming clicks or search efforts between those who posed a negative query and those who posed a positive query. However, when examining the differences in query sentiment across different topics, a notable distinction emerged for participants conducting the search task on the topic of Obesity. Participants who posed a negative query when questioning whether obesity is a disease demonstrated a significantly lower proportion of clicks on attitude-confirming results (see Figure 6.10c). Interestingly, no significant differences were observed between the treatment groups when comparing participants who posed a negative query with those who posed a positive query.



(a) Relation between trait IH and confirmation bias

(b) Knowledge gain

(c) Query Sentiment and Confirmation Bias for participants in topic Obesity

Figure 6.10: Visualization of exploratory analysis on Human Related Factors.

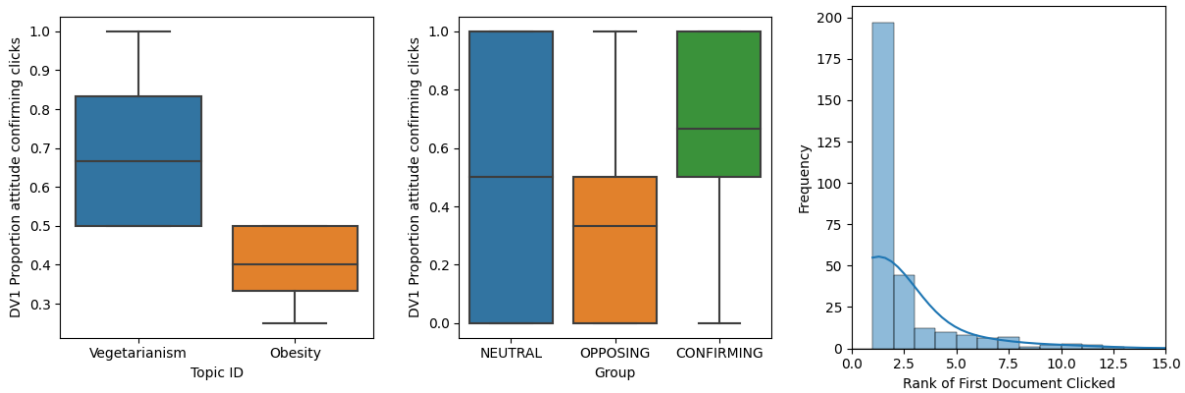
Task Related Factors

We observed various differences between participants assigned to different topics within the same treatment groups. For example, the ratio of attitude-confirming clicks for the topic Vegetarianism was considerably higher compared to the topic Obesity for participants in the Prime group (see Figure 6.11a). While we observed some variations among these subgroups, no apparent pattern emerged.

The analysis of the SERP types and their impact on participants' search behavior revealed interesting findings. Participants in the *Confirming* group demonstrated a significantly higher ratio of clicks on attitude-confirming results. On the other hand, participants in the *Opposing* group exhibited a substantially lower proportion of clicks on attitude-confirming results, while those in the *Neutral* group demonstrated an inconsistent pattern (see Figure 6.11b). These observations prompted further investigation into whether the SERP type influences the rank of the first result clicked and whether participants engage with multiple results.

We observed that only a small subset of participants (13%) displayed a sequential clicking behavior (clicking on the first, second, and third result in order). This finding indicates that a majority of the participants did not show this specific clicking behavior, meaning that there is a higher chance of the first result unconsciously influencing the behavior of the participants. Moreover, we noticed that participants behaved according to expectations in terms of the rank of the first result clicked (see Figure 6.11c). And finally, we did not observe any patterns of the treatment groups influencing the rank of the first result clicked.

The varying search behaviors across groups for the SERP might be due to participants clicking solely on the first document. This would alter the ratio of clicks on attitude-confirming results as the document viewpoints in SERP groups change. To verify this and rule out the possibility of participants clicking only one result, we analyzed the number of clicks made by participants. Cumulative clicks across the SERP groups were relatively consistent, averaging around three. Also, only 17 percent of participants clicked on only one result. Interestingly, the interventions did not seem to significantly alter the behavior of participants within a SERP group in terms of the proportion of the first result clicked.



(a) Topic differences and Confirmation Bias within the Prime Treatment Condition

(b) SERP group and Confirmation Bias

(c) First Click Rank

Figure 6.11: Visualization of exploratory analysis on Task Related Factors.

Part III

Closure

7

Discussion

In this research, we investigated the influence of a boost targeting Intellectual Humility (IH) on individuals' web search behaviors, particularly when engaging with debating topics. We examined whether boosting IH could mitigate confirmation bias, as well as encourage a more comprehensive search effort when searching the web on debated topics. To examine this relationship, we conducted two experiments to create and investigate the impact of IH-boosting interventions. In this chapter, we discuss the findings of these experiments in a broader context, discussing the research fields' implications and limitations.

Moreover, we discuss exploratory analyses, which provide valuable insights into the potential factors influencing the impact of the interventions in boosting IH. It is important to note that further investigation is required to draw definitive conclusions due to the limitations of sample size and the absence of specific hypotheses for the exploratory analyses.

7.1. Study 1, Boosting IH

In Study 1, we conducted a between-subject user study to assess the impact of interventions on State IH. Our aim was to enhance individuals' openness to challenging their beliefs and foster a greater acceptance of potential inaccuracies in their understanding. The interventions included a *trait IH questionnaire*, an *IH prime*, and the *Full Intervention*, each designed to stimulate IH through different mechanisms.

7.1.1. Key findings

In Study 1, we observed that all three interventions successfully enhanced participants' state IH (see Chapter 6). These findings align with previous studies, supporting the notion that IH can be temporarily boosted through simple and cost-effective techniques [34]. Porter et al. [38] also found that merely reading about the benefits of IH could boost self-reported IH. This supports our findings related to the IH prime intervention. Our study builds upon this research by utilizing an IH questionnaire as a means to prompt participants to acknowledge their intellectual fallibility, leading to boosts in IH.

Overall, the outcomes of Study 1 confirm the feasibility and effectiveness of boosting IH through these interventions, providing an encouraging foundation for the next steps of our study. This offers insights into how IH can be increased through targeted interventions and suggest that we might be able to encourage less biased, more open-minded thinking through carefully designed interventions.

7.1.2. Exploratory Findings

In addition to assessing the effectiveness of the interventions, our study conducted exploratory analyses to gain deeper insights into the data. We sought to understand the relationships between participant characteristics and the impact of the interventions, with a specific focus on their initial state and trait IH scores. While no significant correlations were found between initial state IH scores, trait IH scores, and the state IH difference, interesting patterns emerged when considering the strength of participants' opinions. For instance, participants in the group discussing the topic ”Is drinking milk healthy?” exhibited a higher increase in State IH compared to those in the group exploring the topic ”Are Social Networking Sites Good for Our Society?” This suggests that topic-related factors such as emotionality or involvement may have influenced participants' IH levels.

Furthermore, we delved deeper into the variations in viewpoint within each topic. We found that participants in the topics ”Should people become vegetarian?” and ”Should zoos exist?” experienced a significantly greater increase in State IH compared to participants with the opposite viewpoint. Additionally, we explored differences among participants based on their initial opinions. Within the ”Are Social Networking Sites Good for Our Society?” group, for example, participants with strong initial opinions experienced an increase in State IH, while those with less strong opinions exhibited a decrease. This highlights the potential influence of participants' pre-existing attitudes on the effectiveness of the interventions.

7.2. Study 2, Mitigating Confirmation Bias

Encouraged by the findings of the first experiment in Study 1, we conducted our second experiment in Study 2 to test the effects of an IH-boosting intervention on search behavior, which was another between-subject user study. In this study, we expanded the treatment conditions to include control groups and the three interventions. In this research stage, we moved past the assessment of changes in State IH and turned our attention to the practical effects of the interventions on various web search behavior aspects. We measured confirmation bias through the metric *Proportion of attitude-confirming clicks*, and we measured the search effort through the metrics *Lowest Document Rank*, *Average Dwelling Time*, *Task Completion Time*, and *Cumulative Clicks*.

7.2.1. Key findings

In contrast to Study 1, we did not find evidence for our hypotheses of Study 2. We hypothesized that boosting IH would lower biased search behavior, leading to a decrease in the proportion of attitude-confirming clicks and an increased willingness to explore diverse viewpoints (**H3**). In addition, the other hypotheses (**H4-7**) proposed that an enhanced IH would result in more search effort, demonstrated through more document clicks, longer search times, extended reading times, and clicking on lower-ranked search results. Our hypotheses were not confirmed, suggesting that there is no evidence for an impact of the implemented IH boosting interventions on the web search behavior of participants engaging in debates on selected topics. It is crucial to recognize that these findings contribute valuable insights into IH's largely unexplored research territory and its practical implications, particularly in the area of web search behavior. In an era of information overload and online echo chambers, exploring the effects of interventions aimed at fostering IH remains a worthwhile effort. Therefore, these findings should be viewed as an invitation for further research rather than a conclusive endpoint.

7.2.2. Exploratory Findings

We have conducted exploratory analyses to shed light on a variety of variables that we did not account for, which may have added noise to the data and caused the lack of observed differences among treatment groups. These variables are outlined in Section 4.7.

Measurement of search behavior To ensure a thorough assessment of search behavior in this uncharted context, we employed an extensive set of additional metrics. In addition, our analysis of the exploratory search behavior variables indicated no significant differences between the intervention and control groups. Even when comparing the treatment groups (Prime, Questionnaire, and Full intervention) to the control group, no significant differences in search behavior were observed across the exploratory web search behavior measurements. This suggests that other factors, such as the structure of the experiment itself, could be at play in failing to produce evidence of distinct search behaviors across the various groups. Therefore, a more thorough evaluation of the search task's different facets is necessary.

Web search behavior of participants We evaluated whether the observed results may be attributed to a lack of engagement with the search task. However, participants' reflections on the search task indicated that they approached it consistently with their typical search behavior. When asked if they approached the search task similarly to how they usually approach such tasks, over 85% of participants responded affirmatively, indicating that the search task was approached as usual.

We compared our participants' behavior with the behavior reported in [54, 47, 44, 22, 63], given that the design of the task was based on these works. Regarding the rank of the first document clicked, our findings align with the expected pattern, where the first document in the search engine results page (SERP) received

the highest number of clicks, followed by the second document, and so on. The average cumulative clicks per participant were approximately three, which falls between the values reported by Rieger et al. [22] and Draws et al. [63], but lower than those reported by Pothirattanachaikul et al. [52] and Pothirattanachaikul et al. [44]. In terms of dwell time on the SERP, participants spent an average of around 120 seconds, which exceeded the durations reported by Pothirattanachaikul et al. [52] and Suzuki and Yamamoto [50], but fell below the values reported by Draws et al. [63]. The task completion duration averaged around 230 seconds, which was longer than the duration reported by Yamamoto, Yamamoto, and Fujita [100], but shorter than that reported by Pothirattanachaikul et al. [52]. Our findings indicate that participants' behavior during the search task was not significantly divergent from the patterns observed in the literature.

Workload: NASA-TLX Our exploration into the perceived workload and search effort revealed a complex relationship, suggesting that cognitive load and search behaviors might be influenced by a variety of factors. This intricate relationship was further highlighted when considering the influence of specific topics. Notably, we found that different topics impacted participants' perceived workload in diverse ways, emphasizing the role of topic-specific factors in shaping the cognitive load during the search process. This suggests that certain topics may inherently require more cognitive effort, possibly due to their inherent complexity, contentious nature, or the level of pre-existing knowledge required.

Moreover, our findings indicated an intriguing relationship between trait IH scores and perceived workload. Participants scoring higher in trait IH tended to report a lower workload, which aligns with the notion that individuals with higher levels of IH may approach tasks with a greater sense of ease and accomplishment, as suggested by Porter et al. [38]. Interestingly, these participants, with higher trait IH scores, also demonstrated higher NASA Performance and NASA Effort scores, suggesting a sense of achievement and increased effort in their search activities.

Demographics Age appeared to play a role in shaping participants' interactions with search results. We discovered that specific age groups exhibited a decreased inclination to click on attitude-confirming results, implying a greater openness to diverse viewpoints within these demographics. In particular, participants aged 18-20 demonstrated a notably longer average dwelling duration, suggesting a more thorough engagement with the information, reflecting more effortful search behavior. Considering these findings, it becomes clear that varying strategies might be required to effectively influence search behaviors across distinct age demographics.

Occupation, specifically being a student, was another influential factor. We found that students demonstrated a decreased inclination to click on attitude-confirming results and a more thorough and diligent approach to their searches. This behavior could potentially be attributed to academic conditioning that encourages comprehensive research and critical thinking. These findings are consistent with the research conducted by Zou et al. [103], which highlighted differences in search behavior based on educational background. Remarkably, students assigned to the topic of School Uniforms exhibited a relatively low proportion of clicks on attitude-confirming results compared to the other topics, hinting at a more neutral or open attitude towards this particular topic among students. Contrasting these findings, we did not find any significant disparities in search behavior between genders, unlike the findings reported by Kim, Lehto, and Morrison [107].

Human Centered Factors A key observation is the lower proportion of attitude-confirming clicks among participants possessing higher trait IH, echoing previous work by Porter and Schumann [37] that identified a similar negative correlation. Conversely, our results show no discernible connection between trait IH levels and search effort, implying that while trait IH may influence the propensity to click on attitude-confirming results, it may not necessarily impact the overall search effort exerted by participants.

Further exploration revealed an unexpected relationship between knowledge gain and confirmation bias, with participants reporting no knowledge gain demonstrating the lowest proportion of clicks on attitude-confirming results. Yet, it appeared that the interventions employed in the study did not impact participants' self-reported acquisition of topic-specific knowledge, as no significant differences in self-reported knowledge gain across different treatment conditions were found.

When examining viewpoint intensity, we found that neither participant intensity nor the interventions deployed significantly impacted search behavior. This suggests that these factors may not play a dominant

role in shaping individuals' information search strategies.

Finally, concerning query sentiment, we found no significant differences in the proportion of attitude-confirming clicks or search efforts, indicating that the emotional tone of a query may not necessarily affect the balance of information that participants seek. An intriguing exception, however, was observed in the context of the Obesity topic, where participants posing a negative query demonstrated a significantly lower proportion of clicks on attitude-confirming results, suggesting that, at least for certain topics, query sentiment may influence participants' tendency to click on attitude-confirming results.

Task Related Factors In the realm of search engine result pages (SERPs), the behaviors of our study participants revealed intriguing patterns. Participants in the *Confirming* group demonstrated a significantly higher ratio of clicks on attitude-confirming results, aligning with the outcomes of Pothirattanachaikul et al. [52], which suggests that search engines can influence users' polarization by presenting belief-consistent results at higher ranks.

An interesting finding was that only a small subset of participants (13%) displayed a sequential clicking behavior (clicking on the first, second, and third result in order). This suggests that a majority of the participants did not default to this particular behavior, lending support to the notion that the first search result may hold an unconscious influence over user behavior.

When exploring cumulative clicks across the SERP groups, we found them to be relatively consistent, in line with the findings of Rieger et al. [22] and Draws et al. [63], averaging around three. This observation, coupled with the fact that a majority of participants, irrespective of the SERP group, clicked on more than one document, indicates a propensity among participants to explore multiple results before concluding their search task. This inclination towards multiple perspectives contrasts with the findings of Pothirattanachaikul et al. [44], who reported a decreased likelihood of SERP interaction when participants encounter a belief-inconsistent answer at the outset.

These findings collectively underscore the potential influence of the SERP group on participant behavior in relation to confirmation bias, possibly exerting a more substantial impact than the interventions we employed.

7.3. Limitations

Despite the rigorous, preregistered methodology, both user studies have several limitations that should be considered when interpreting the results. First, there was a limited measurement of IH in Study 2. Although we measured it for participants in the *Questionnaire* and *Full intervention* conditions, we did not assess trait IH for all participants after the experiment, given its irrelevance to our hypotheses. Nonetheless, our findings in these conditions indicated that participants with lower trait IH experienced greater workload and exhibited a negative correlation between confirmation bias and trait IH, aligning with [34]. Future research should consider measuring trait IH for all participants and examining pre- and post-test measurements of state IH to gain additional insights.

The constrained sample sizes utilized in the exploratory analyses of both studies were a result of focusing on specific participant subgroups, such as individuals with certain characteristics. The detailed focus on these specialized groups limited the overall sample size, potentially influencing the results and constraining the broad applicability of the findings. To confirm and expand upon the initial insights gathered from this exploratory analysis, further research necessitates the use of larger sample sizes and more concentrated hypotheses.

Another limitation relates to the limited search environment employed in Study 2. Although the search task was argued to be engaging, a more free-form search task might have yielded richer data and presented a more realistic search scenario. The lack of flexibility in the search tasks, where the search results were generated from a fixed dataset unaffected by participants' queries, limited the diversity of search behaviors. While introducing a more free-form search experiment could potentially provide more nuanced insights, it also poses challenges in terms of replicability. Thus, we initially chose a controlled setting to establish the efficacy of our intervention within a specific context. Future research could consider implementing less constrained search tasks, or task variations, to further examine the outcomes and implications of our findings. Addressing these limitations can contribute to a more robust understanding of the relationship between IH-boosting interventions and web search behavior.

8

Conclusion

In this master's thesis, we pioneered the investigation of whether boosting Intellectual Humility (IH) is effective in mitigating confirmation bias during web search on debated topics. The research encompassed two distinct studies: one focused on testing the efficacy of three IH-boosting interventions, and the other aimed to test these interventions within the context of web search behavior. By addressing the research objectives and research questions, our investigation has made significant contributions to the understanding of IH and its impact on web search behavior.

The overarching research questions that guided both studies were:

Research Question 1

Does boosting participants' IH through priming, a trait IH questionnaire, or a combination of both interventions affect their level of IH?

Research Question 2

What is the impact of a boosting application on confirmation bias during web search?

Research Question 3

Can confirmation bias during Web Search be mitigated by boosting Intellectual Humility?

Method

Study 1 aimed to determine the effectiveness of the interventions in increasing participants' IH levels. Through the implementation of the *trait IH questionnaire*, *IH prime*, and *full* interventions, we successfully assessed the impact of these interventions on state IH. Study 2 focused on evaluating the effects of these interventions on web search behavior. By examining participants' search behaviors post-intervention compared to a control group, we aimed to extend the literature by understanding the influence of boosting IH on search behavior and by focusing on the internal user-centric challenges rather than the external challenges in the digital environment. The analysis of the data and the comparison of treatment groups with the control group provided valuable insights into the shaping of participants' information-seeking strategies.

Results

Reflecting on the research objectives and research questions, we can conclude that our study has successfully addressed these key areas. We have boosted participants' IH through priming, a trait IH questionnaire, and a combination of both interventions. And we have examined the relationship between boosting IH and search behavior, where we did not observe substantial differences between the treatment groups and the control group. Nevertheless, we explored the role of participant characteristics and topic characteristics in the effectiveness of the interventions and observed that both individual and environmental

factors, including education level, personal viewpoints, and search results order, shape the impact of IH-boosting interventions on online search behavior, with varying effects observed across different debated topics. The findings shed light on the complexities of modifying online information-seeking behaviors and emphasize the importance of promoting unbiased and responsible information retrieval and evaluation.

Contributions

The contributions of this research extend considerably beyond mere theoretical significance. A pivotal part of this work is the exploration, comparison, and provision of empirical evidence for IH-boosting interventions. Our findings demonstrate the potential of these interventions to effect significant increases in state IH. This sets the foundation for future studies and practical applications, offering valuable insights that can guide the design of similar interventions aimed at boosting IH.

This research offers a nuanced understanding of the complex relationship between various IH-boosting interventions, the mitigation of confirmation bias, and behavior in web searches by introducing a novel intervention approach. Through an empirical analysis, we have identified the complex interplay of these factors. This relationship analysis holds promising implications for both individuals and society. It contributes to the development of practical solutions that enable thorough and unbiased searches on debated topics, which promotes responsible opinion formation and decision-making, improving the quality of information-seeking practices in online environments.

Also, we contribute a rich dataset containing behavioral data from the two studies conducted. Firstly, it contains information on the effect of three different boosting interventions on state IH. Secondly, it contains data from search logs, measures of knowledge, attitude, mood, emotion, receptiveness to opposing views, and demographics, indicating the effect of the boosting interventions on confirmation bias and on web search behavior. Put together, this dataset contains information from more than 500 users. The dataset serves as a rich resource for future research, offering deep insights into the practical application and effects of IH-boosting interventions.

Yet another contribution consists of the sharing of the pre-registrations of the user studies conducted, enabling other researchers to replicate these studies and verify our findings. And finally, a key contribution is the search interface software project. This tool can be conveniently modified for various use cases, particularly for experiments requiring the monitoring of search behavior. As the exploratory findings have unveiled numerous compelling avenues for future research, this tool could be employed by other researchers to examine these fascinating possibilities.

Reflection

In reflecting on the methodology and research process employed in this research, the chosen research design proved appropriate in addressing the research objectives. The intervention approach, data collection techniques, and comprehensive analysis of subjective and objective measures allowed for a thorough examination of IH-boosts and their impact on web search behavior on debated topics. However, it is important to acknowledge the strengths and limitations of our data collection and analysis methods. The use of self-report scales and log data provided a comprehensive understanding of participants' attitudes, search behavior, and their interactions with the search engine. These measures allowed us to capture both subjective and objective aspects of their information-seeking process. Additionally, the rich exploratory data expanded the scope of our analysis, providing a more comprehensive picture of web search behavior in the investigated context. Nevertheless, the sample size constraints and the absence of specific hypotheses for exploratory analyses should be taken into consideration when interpreting the exploratory findings.

Future Work

Moving forward, future research endeavors should strive to expand on the findings of this study while addressing its identified limitations. One promising avenue to pursue involves exploring alternative IH-boosting approaches. This could involve developing new interventions (potentially based on the different boost types as described in Table 2.1), comparing different types of interventions, and investigating the effects of tailored interventions. This would add to our understanding of the range of tools available for mitigating confirmation bias and enhancing information-seeking practices.

A detailed examination of the separate components of IH and their respective impacts on information

behavior is another area for potential future explorations. By isolating the different aspects of IH, we could gain a more nuanced understanding of how each component interacts with web search behavior and confirmation bias. It would also be interesting for future studies to measure IH for all participants, not only to gain a thorough understanding of the influence of initial levels of IH on the effectiveness of different interventions but also to capture potential variations among individuals. This could help tailor interventions more effectively. Furthermore, future research could incorporate more realistic search environments to capture the effects of IH-boosting interventions in real-world settings. This could include the use of real search engines and exploring the effects of IH-boosting interventions in real-life decision-making scenarios. Studying the long-term effects of sustained or repeated IH-boosting interventions would also help us understand the durability of the intervention effects and whether they can be amplified over time.

In conclusion, this master's thesis has demonstrated the potential of boosting IH to influence web search behavior on debated topics and mitigate confirmation bias. This thesis demonstrates that boosting can be utilized to mitigate confirmation bias, shifting away from nudging and laying the foundation for further research on boosting and biases. Our findings hint at a need for a nuanced, personalized approach in the design of these interventions aimed at fostering IH. It is our hope that this study inspires and is an initial basis for continued efforts in exploring the multifaceted relationship between IH, information-seeking behavior, and responsible opinion formation, ultimately promoting a more informed and unbiased online discourse.

References

[1] Noel Carroll. “In search we trust: exploring how search engines are shaping society”. In: *International Journal of Knowledge Society Research (IJKSR)* 5.1 (2014), pp. 12–27.

[2] Ingmar Weber et al. “Who uses web search for what: and how”. In: *Proceedings of the fourth ACM international conference on Web search and data mining*. 2011, pp. 15–24.

[3] Alisa Rieger et al. “Searching for the Whole Truth: Harnessing the Power of Intellectual Humility to Boost Better Search on Debated Topics”. In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–8.

[4] Philipp Lorenz-Spreen et al. “How behavioural sciences can promote truth, autonomy and democratic discourse online”. In: *Nature Human Behaviour* 4 (11 Nov. 2020), pp. 1102–1109. DOI: 10.1038/s41562-020-0889-7.

[5] Alisa Rieger. “Interactive Interventions to Mitigate Cognitive Bias”. In: Association for Computing Machinery, Inc, Apr. 2022, pp. 316–320. DOI: 10.1145/3503252.3534362.

[6] Philipp Lorenz-Spreen et al. “Boosting people’s ability to detect microtargeted advertising”. In: *Scientific Reports* 11 (1 Dec. 2021). DOI: 10.1038/s41598-021-94796-z.

[7] M F Beijen. *Effects of Time Constraints and Search Results Presentation on Web Search*.

[8] Varol Onur Kayhan. “Confirmation Bias: Roles of Search Engines and Search Contexts”. In: (2015).

[9] Daniel T Gilbert. “How mental systems believe.” In: *American psychologist* 46.2 (1991), p. 107.

[10] Thomas T Hills. “The dark side of information proliferation”. In: *Perspectives on Psychological Science* 14.3 (2019), pp. 323–330.

[11] Raymond S Nickerson. *Confirmation Bias: A Ubiquitous Phenomenon in Many Guises*. 1998, pp. 175–220.

[12] Eric Rassin. *Individual differences in the susceptibility to confirmation bias*. 2008, pp. 87–93.

[13] Charles G Lord et al. “Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence.” In: *Journal of personality and social psychology* 37.11 (1979), p. 2098.

[14] Armen E Allahverdyan et al. “Opinion dynamics with confirmation bias”. In: *PloS one* 9.7 (2014), e99557.

[15] Loren J Chapman. “Illusory correlation in observational report”. In: *Journal of Verbal Learning and Verbal Behavior* 6.1 (1967), pp. 151–155.

[16] Gloria Phillips-Wren et al. “Decision making under stress: The role of information overload, time pressure, complexity, and uncertainty”. In: *Journal of Decision Systems* 29.sup1 (2020), pp. 213–225.

[17] Daniel Kahneman et al. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press, 1982.

[18] Scott O Lilienfeld et al. “Giving debiasing away: Can psychological research on correcting cognitive errors promote human welfare?” In: *Perspectives on psychological science* 4.4 (2009), pp. 390–398.

[19] A. Sellier et al. *Debiasing Training Improves Decision Making in the Field*. June 2020. DOI: 10.1177/0956797620930211.

[20] Norah E. Dunbar et al. “Implicit and explicit training in the mitigation of cognitive bias through the use of a serious game”. In: *Computers in Human Behavior* 37 (2014), pp. 307–318. DOI: 10.1016/j.chb.2014.04.053.

[21] Jackie M. Poos et al. “Battling bias: Effects of training and training context”. In: *Computers and Education* 111 (Aug. 2017), pp. 101–113. DOI: 10.1016/j.compedu.2017.04.004.

[22] Alisa Rieger et al. “This Item Might Reinforce Your Opinion: Obfuscation and Labeling of Search Results to Mitigate Confirmation Bias”. In: Association for Computing Machinery, Inc, Aug. 2021, pp. 189–199. DOI: 10.1145/3465336.3475101.

[23] Ralph Hertwig et al. “Nudging and Boosting: Steering or Empowering Good Decisions”. In: *Perspectives on Psychological Science* 12 (6 Nov. 2017), pp. 973–986. DOI: 10.1177/1745691617702496.

[24] Ana Caraban et al. “23 Ways to Nudge: A review of technology-mediated nudging in human-computer interaction”. In: Association for Computing Machinery, May 2019. DOI: 10.1145/3290605.3300733.

[25] Anastasia Kozyreva et al. “Citizens Versus the Internet: Confronting Digital Challenges With Cognitive Tools”. In: *Psychological Science in the Public Interest* 21 (3 Dec. 2020), pp. 103–156. DOI: 10.1177/1529100620946707.

[26] Jon Roozenbeek et al. “The fake news game: actively inoculating against the risk of misinformation”. In: *Journal of Risk Research* 22 (5 May 2019), pp. 570–580. DOI: 10.1080/13669877.2018.1443491.

[27] Sander van der Linden et al. “Inoculating the Public against Misinformation about Climate Change”. In: *Global Challenges* 1 (2 Feb. 2017), p. 1600008. DOI: 10.1002/gch2.201600008.

[28] Leif Azzopardi. “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*. 2021, pp. 27–37.

[29] Ryen W White et al. “Belief dynamics and biases in web search”. In: *ACM Transactions on Information Systems (TOIS)* 33.4 (2015), pp. 1–46.

[30] Rick H Hoyle et al. “Holding specific views with humility: Conceptualization and measurement of specific intellectual humility”. In: *Personality and Individual Differences* 97 (2016), pp. 165–172.

[31] Mark R. Leary et al. “Cognitive and Interpersonal Features of Intellectual Humility”. In: *Personality and Social Psychology Bulletin* 43 (6 June 2017), pp. 793–813. DOI: 10.1177/0146167217697695.

[32] Shauna M Bowes et al. “Stepping outside the echo chamber: Is intellectual humility associated with less political myside bias?” In: *Personality and Social Psychology Bulletin* 48.1 (2022), pp. 150–164.

[33] Leor Zmigrod et al. “The psychological roots of intellectual humility: The role of intelligence and cognitive flexibility”. In: *Personality and Individual Differences* 141 (Apr. 2019), pp. 200–208. DOI: 10.1016/j.paid.2019.01.016.

[34] Tenelle Porter et al. “Predictors and consequences of intellectual humility”. In: *Nature Reviews Psychology* 1.9 (2022), pp. 524–536.

[35] Mark R Leary. *Intellectual Humility as a Route to More Accurate Knowledge, Better Decisions, and Less Conflict*. 2022.

[36] Rink Hoekstra et al. “Aspiring to greater intellectual humility in science”. In: *Nature human behaviour* 5.12 (2021), pp. 1602–1607.

[37] Tenelle Porter et al. “Intellectual humility and openness to the opposing view”. In: *Self and Identity* 17.2 (2018), pp. 139–162.

[38] Tenelle Porter et al. “Intellectual humility predicts mastery behaviors when learning”. In: *Learning and Individual Differences* 80 (2020), p. 101888.

[39] Don E Davis et al. “Distinguishing intellectual humility and general humility”. In: *The Journal of Positive Psychology* 11.3 (2016), pp. 215–224.

[40] Tim Draws et al. “This Is Not What We Ordered: Exploring Why Biased Search Result Rankings Affect User Attitudes on Debated Topics”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021, pp. 295–305.

[41] Carol C Kuhlthau. “Inside the search process: Information seeking from the user’s perspective”. In: *Journal of the American society for information science* 42.5 (1991), pp. 361–371.

[42] Leif Azzopardi. “Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval”. In: Association for Computing Machinery, Inc, Mar. 2021, pp. 27–37. DOI: 10.1145/3406522.3446023.

[43] Marwah Alaofi et al. “Where Do Queries Come From?” In: Association for Computing Machinery, Inc, July 2022, pp. 2850–2862. DOI: 10.1145/3477495.3531711.

[44] Suppanut Pothirattanachaikul et al. “Analyzing the effects of ”people also ask” on search behaviors and beliefs”. In: Association for Computing Machinery, Inc, July 2020, pp. 101–110. DOI: 10.1145/3372923.3404786.

[45] Suppanut Pothirattanachaikul et al. “Analyzing the effects of document’s opinion and credibility on search behaviors and belief dynamics”. In: Association for Computing Machinery, Nov. 2019, pp. 1653–1662. DOI: 10.1145/3357384.3357886.

[46] Leif Azzopardi et al. “How query cost affects search behavior”. In: *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 2013, pp. 23–32.

[47] Dáša Vedejová et al. “Confirmation bias in information search, interpretation, and memory recall: evidence from reasoning about four controversial topics”. In: *Thinking and Reasoning* 28 (1 2022), pp. 1–28. DOI: 10.1080/13546783.2021.1891967.

[48] Ryen W. White et al. “Belief dynamics and biases in web search”. In: *ACM Transactions on Information Systems* 33 (4 Apr. 2015). DOI: 10.1145/2746229.

[49] Stefan Schweiger et al. “Confirmation bias in web-based search: A randomized online study on the effects of expert information and social tags on information search and evaluation”. In: *Journal of Medical Internet Research* 16 (3 2014). DOI: 10.2196/jmir.3044.

[50] Masaki Suzuki et al. “Characterizing the Influence of Confirmation Bias on Web Search Behavior”. In: *Frontiers in Psychology* 12 (2021), p. 771948.

[51] Yusuke Yamamoto et al. “Query priming for promoting critical thinking in web search”. In: *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. 2018, pp. 12–21.

[52] Suppanut Pothirattanachaikul et al. “Analyzing the effects of document’s opinion and credibility on search behaviors and belief dynamics”. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2019, pp. 1653–1662.

[53] Gary Marchionini. “Exploratory search: from finding to understanding”. In: *Communications of the ACM* 49.4 (2006), pp. 41–46.

[54] Kumaripaba Athukorala et al. “Is exploratory search different? A comparison of information search behavior for exploratory and lookup tasks”. In: *Journal of the Association for Information Science and Technology* 67.11 (2016), pp. 2635–2651.

[55] Benjamin A Clegg et al. “Game-based training to mitigate three forms of cognitive bias”. In: *Proceedings of Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*. Vol. 14180. 2014, pp. 1–12.

[56] Calum Thornhill et al. “A Digital Nudge to Counter Confirmation Bias”. In: *Frontiers in Big Data* 2 (June 2019). DOI: 10.3389/fdata.2019.00011.

[57] Stefan Schweiger et al. “Confirmation bias in web-based search: a randomized online study on the effects of expert information and social tags on information search and evaluation”. In: *Journal of medical Internet research* 16.3 (2014), e94.

[58] Hsieh Hong Huang et al. “How to better reduce confirmation bias? The fit between types of counter-argument and tasks”. In: (2010).

[59] Andrew White. “Overcoming ‘confirmation bias’ and the persistence of conspiratorial types of thinking”. In: *Continuum* 36.3 (2022), pp. 364–376.

[60] Varol Onur Kayhan. “Seeking health information on the web: positive hypothesis testing”. In: *International journal of medical informatics* 82.4 (2013), pp. 268–275.

[61] Baruch Fischoff. *Debiasing*. Tech. rep. Decision Research Eugene OR, 1981.

[62] J. E. Korteling et al. *Retention and Transfer of Cognitive Bias Mitigation Interventions: A Systematic Literature Study*. Aug. 2021. DOI: 10.3389/fpsyg.2021.629354.

[63] Tim Draws et al. “This Is Not What We Ordered: Exploring Why Biased Search Result Rankings Affect User Attitudes on Debated Topics”. In: Association for Computing Machinery, Inc, July 2021, pp. 295–305. DOI: 10.1145/3404835.3462851.

[64] Hsieh Hong Huang et al. *How to better reduce confirmation bias? The fit between types of counter-argument and tasks.* 2010. URL: http://aisel.aisnet.org/sais2010/7.

[65] Hsieh Hong Huang et al. “Understanding the role of computer-mediated counter-argument in countering confirmation bias”. In: *Decision Support Systems* 53 (3 June 2012), pp. 438–447. DOI: 10.1016/j.dss.2012.03.009.

[66] Derek J Koehler et al. *Blackwell handbook of judgment and decision making*. John Wiley & Sons, 2008.

[67] Pelle Guldborg Hansen et al. “Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy”. In: *European Journal of Risk Regulation* 4.1 (2013), pp. 3–28.

[68] Jessica E Brodsky et al. “Improving college students’ fact-checking strategies through lateral reading instruction in a general education civics course”. In: *Cognitive research: principles and implications* 6.1 (2021), pp. 1–18.

[69] Melisa Basol et al. “Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news”. In: *Journal of Cognition* 3 (1 2020). DOI: 10.5334/joc.91.

[70] John Cook et al. “The cranky uncle game—Combining humor and gamification to build student resilience against climate misinformation”. In: *Environmental Education Research* (2022), pp. 1–17.

[71] Nabil F Saleh et al. “Active inoculation boosts attitudinal resistance against extremist persuasion techniques: A novel approach towards the prevention of violent extremism”. In: *Behavioural Public Policy* (2021), pp. 1–24.

[72] S Wineburg. “To navigate the dangers of the web, you need critical thinking–but also critical ignoring”. In: *The Conversation. https://theconversation. com/to-navigate-the-dangers-of-the-web-you-need-critical-thinking-but-also-critical-ignoring-158617* (2021).

[73] Nicholas H Lurie et al. “Simple decision aids and consumer decision making”. In: *Journal of Retailing* 90.4 (2014), pp. 511–523.

[74] Melisa Basol et al. “Towards psychological herd immunity: Cross-cultural evidence for two pre-bunking interventions against COVID-19 misinformation”. In: *Big Data and Society* 8 (1 2021). DOI: 10.1177/20539517211013868.

[75] Matthew M Martin et al. “A new measure of cognitive flexibility”. In: *Psychological reports* 76.2 (1995), pp. 623–626.

[76] Maggie E. Toplak et al. “Assessing miserly information processing: An expansion of the Cognitive Reflection Test”. In: *Thinking and Reasoning* 20 (2 Apr. 2014), pp. 147–168. DOI: 10.1080/13546783.2013.844729.

[77] Arne Roets et al. “Item selection and validation of a brief, 15-item version of the Need for Closure Scale”. In: *Personality and Individual Differences* 50 (1 Jan. 2011), pp. 90–94. DOI: 10.1016/j.paid.2010.09.004.

[78] Takehiro Yamamoto et al. “Exploring people’s attitudes and behaviors toward careful information seeking in web search”. In: Association for Computing Machinery, Oct. 2018, pp. 963–972. DOI: 10.1145/3269206.3271799.

[79] Matthew L Stanley et al. “Intellectual humility and perceptions of political opponents”. In: *Journal of Personality* 88.6 (2020), pp. 1196–1216.

[80] Jian Du et al. “Owning One’s Intellectual Limitations: A Review of Intellectual Humility”. In: *Psychology* 11.07 (2020), p. 1009.

[81] Elizabeth J Krumrei-Mancuso et al. “Intellectual humility in the sociopolitical domain”. In: *Self and Identity* 19.8 (2020), pp. 989–1016.

[82] Diane Kelly. “Methods for evaluating interactive information retrieval systems with users”. In: *Foundations and Trends in Information Retrieval* 3 (1-2 2009), pp. 1–224. DOI: 10.1561/1500000012.

[83] Benjamin R Meagher et al. “Contrasting self-report and consensus ratings of intellectual humility and arrogance”. In: *Journal of Research in Personality* 58 (2015), pp. 35–45.

[84] Mark Alfano et al. “Development and validation of a multi-dimensional measure of intellectual humility”. In: *PloS one* 12.8 (2017), e0182950.

[85] Sergiu Chelaru et al. “Analyzing, detecting, and exploiting sentiment in web queries”. In: *ACM Transactions on the Web (TWEB)* 8.1 (2013), pp. 1–28.

[86] Marwah Alaofi et al. “Where Do Queries Come From?” In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2022, pp. 2850–2862.

[87] Tim Draws et al. “A checklist to combat cognitive biases in crowdsourcing”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 9. 2021, pp. 48–59.

[88] Martin Potthast et al. “Argument search: Assessing argument relevance”. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2019, pp. 1117–1120.

[89] Gizem Gezici et al. “Evaluation metrics for measuring bias in search engine results”. In: *Information Retrieval Journal* 24 (2021), pp. 85–113.

[90] Samantha A. Deffler et al. “Knowing What You Know: Intellectual Humility and Judgments of Recognition Memory”. In: *Personality and Individual Differences* 96 (2016), pp. 255–259. DOI: 10.1016/j.paid.2016.03.016.

[91] Elizabeth J Krumrei-Mancuso et al. “Links between intellectual humility and acquiring knowledge”. In: *The Journal of Positive Psychology* 15.2 (2020), pp. 155–170.

[92] Tim Gorichanaz. “Relating information seeking and use to intellectual humility”. In: *Journal of the Association for Information Science and Technology* 73.5 (2022), pp. 643–654.

[93] Abigail Gertner et al. *The Assessment of Biases in Cognition Development and Evaluation of an Assessment Instrument for the Measurement of Cognitive Bias*.

[94] P. C. Wason. “On the Failure to Eliminate Hypotheses in a Conceptual Task”. In: *Quarterly Journal of Experimental Psychology* 12 (3 July 1960), pp. 129–140. DOI: 10.1080/17470216008416717.

[95] Mark Snyder et al. *Hypothesis-Testing Processes in Social Interaction*. 1978, pp. 1202–1212.

[96] Vincent Berthet. “The Measurement of Individual Differences in Cognitive Biases: A Review and Improvement”. In: *Frontiers in Psychology* 12 (Feb. 2021). DOI: 10.3389/fpsyg.2021.630177.

[97] Ryen White. “Beliefs and biases in web search”. In: *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 2013, pp. 3–12.

[98] Luyan Xu et al. “How do user opinions influence their interaction with web search results?” In: *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 2021, pp. 240–244.

[99] Samantha A Deffler et al. “Knowing what you know: Intellectual humility and judgments of recognition memory”. In: *Personality and Individual Differences* 96 (2016), pp. 255–259.

[100] Takehiro Yamamoto et al. “Exploring people’s attitudes and behaviors toward careful information seeking in web search”. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 2018, pp. 963–972.

[101] Sandra G. Hart et al. “Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research”. In: *Advances in Psychology*. Ed. by Peter A. Hancock et al. Vol. 52. Human Mental Workload. North-Holland, 1988, pp. 139–183. DOI: 10.1016/S0166-4115(08)62386-9.

[102] Jeonghyun Kim. “Task difficulty as a predictor and indicator of web searching interaction”. In: *CHI’06 extended abstracts on human factors in computing systems*. 2006, pp. 959–964.

[103] Jie Zou et al. “Users meet clarifying questions: Toward a better understanding of user interactions for search clarification”. In: *ACM Transactions on Information Systems* 41.1 (2023), pp. 1–25.

[104] Stefano Baccianella et al. “Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.” In: *Lrec*. Vol. 10. 2010. 2010, pp. 2200–2204.

[105] Suzanne Tolmeijer et al. “Second Chance for a First Impression? Trust Development in Intelligent System Interaction”. In: *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. UMAP ’21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 77–87. DOI: 10.1145/3450613.3456817.

[106] Thomas Franke et al. “A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale”. In: *International Journal of Human–Computer Interaction* 35 (2019), pp. 456–467. DOI: 10.1080/10447318.2018.1456150.

[107] Dae-Young Kim et al. “Gender differences in online travel information search: Implications for marketing communications on the internet”. In: *Tourism management* 28.2 (2007), pp. 423–433.

A

Questionnaires

Items from the state IH questionnaire The items in this questionnaire are taken from Hoyle et al. [30]. In the survey, TOPIC is replaced with the topic on which the participant had indicated to have the strongest opinion.

1. My views about *TOPIC* are just as likely to be wrong as other views.
2. I recognize that my views about *TOPIC* are based on limited evidence.
3. Although I have particular views about *TOPIC*, I realize that I don't know everything that I need to know about it.
4. It is quite likely that there are gaps in my understanding about *TOPIC*.
5. My sources for information about *TOPIC* might not be the best.
6. I am open to new information in the area of *TOPIC* that might change my view.
7. My views about *TOPIC* today may someday turn out to be wrong.
8. When it comes to my views about *TOPIC* I may be overlooking evidence.
9. My views about *TOPIC* may change with additional evidence or information.

Items from the trait IH questionnaire The items in this questionnaire are taken from Alfano et al. [84].

1. I think that paying attention to people who disagree with me is a waste of time.
2. I feel no shame learning from someone who knows more than me.
3. If I do not know much about some topic, I don't mind being taught about it, even if I know about other topics.
4. Even when I have high status, I don't mind learning from others who have lower status.
5. Only wimps admit that they've made mistakes
6. I don't take people seriously if they're very different from me.
7. Being smarter than other people is not especially important to me.
8. I would like to be seen explaining ideas that no one else understands.
9. I get a lot of pleasure from knowing more than other people.
10. I want people to know that I am an unusually intelligent person.
11. I like to be the smartest person in the room.
12. I find it annoying to be told that I've made an intellectual mistake.
13. If someone points out an intellectual mistake that I've made, I tend to get angry.
14. I appreciate being corrected when I make a mistake.
15. When someone corrects a mistake that I've made, I do not feel embarrassed.
16. When I realize that someone knows more than me, I feel frustrated and humiliated.
17. I rarely discuss things that I wish I understood better with other people.

18. I enjoy reading about the ideas of different cultures.
19. I would be very bored by a book about ideas I disagreed with.
20. I've never really enjoyed figuring out why people disagree with me.
21. I find it boring to discuss things I don't already understand.
22. A disagreement is like a war.

IH Prime ***About Intellectual Humility***

Intellectual Humility (IH) means acknowledging one's limitation of knowledge and beliefs, recognizing that these beliefs may be biased, and being open to learning from others and considering alternative viewpoints. IH enhances knowledge, understanding, and the quality of people's decisions.

Intellectually humble people:

- Are more curious
- Are better liked as leaders
- Tend to make more thorough, well-informed decisions
- Seem to be more open to cooperating with those whose views differ from their own

Tip!

To become more intellectually humble, try actively seeking out and engaging with ideas that challenge your own beliefs.

B

Search Engine Visuals

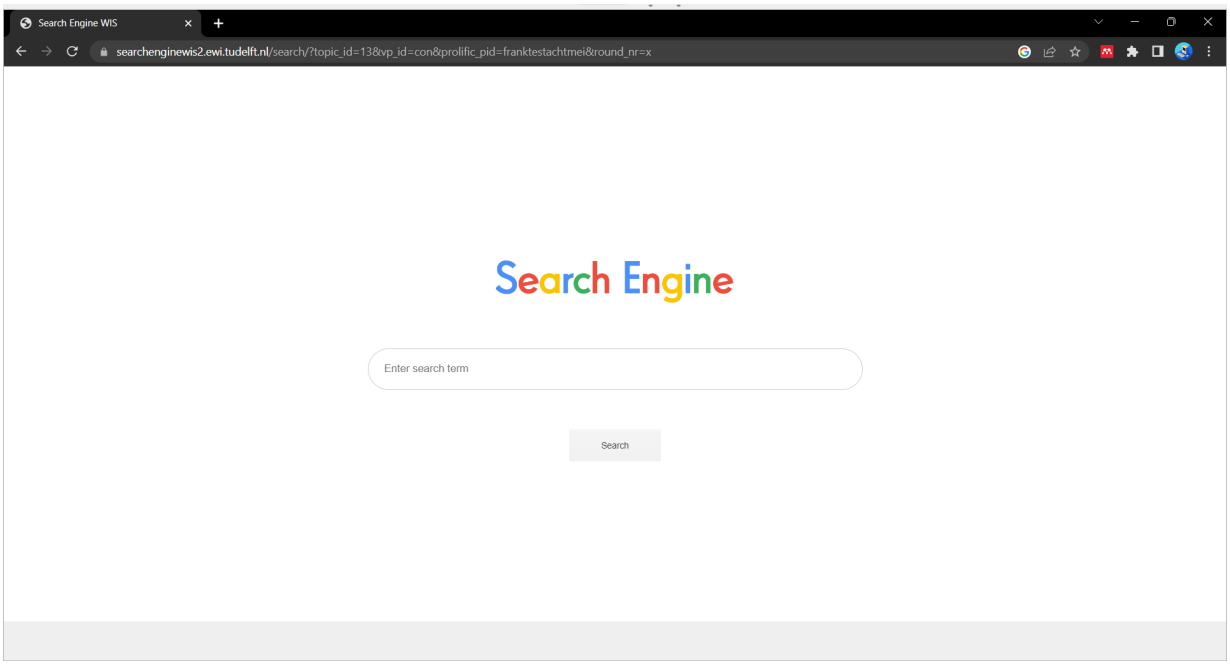


Figure B.1: Screenshot of the search engine upon opening the application.

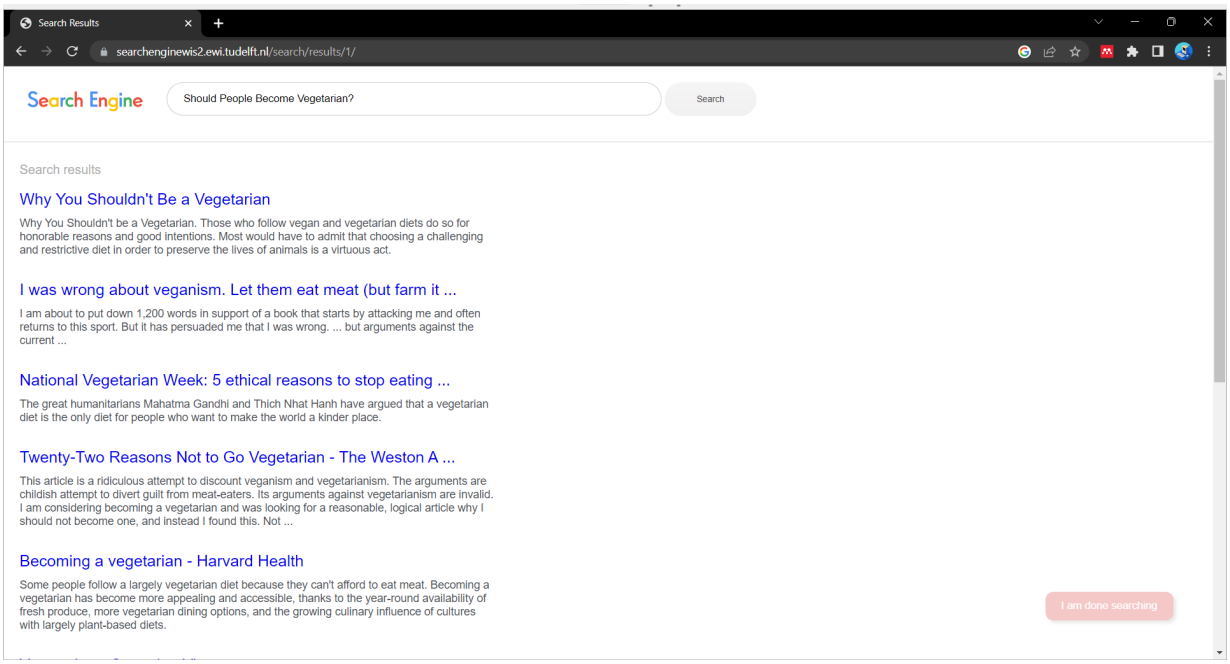


Figure B.2: Screenshot of the search engine results page.

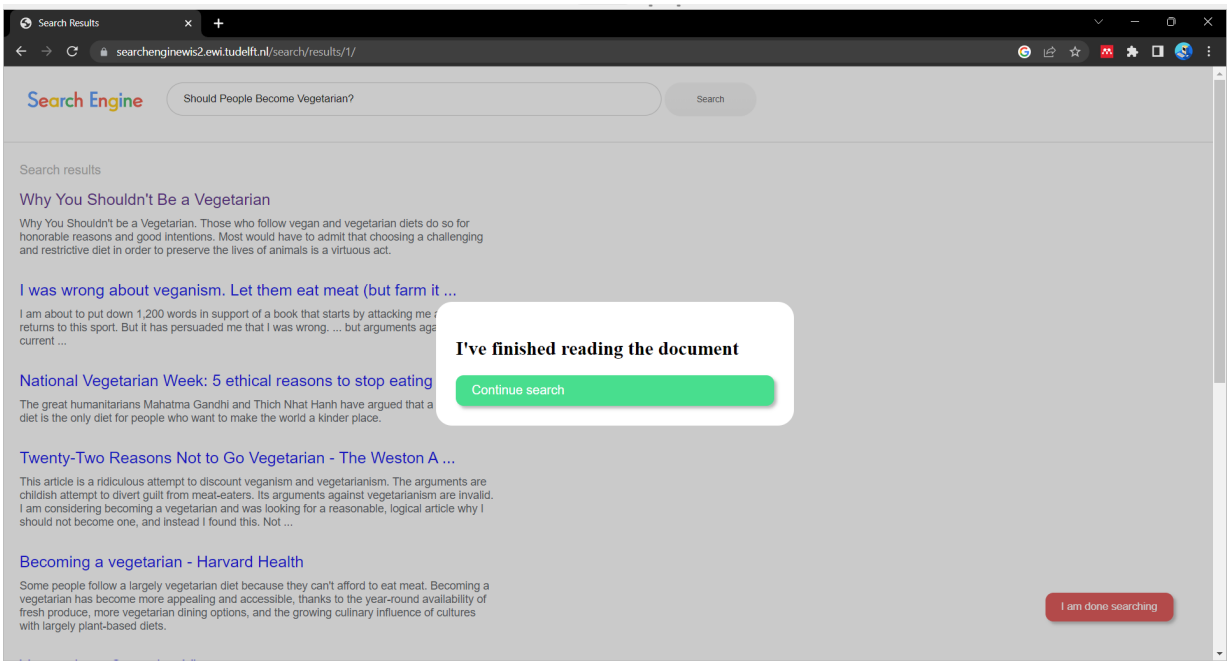


Figure B.3: Screenshot of the search engine after finishing reading a document.

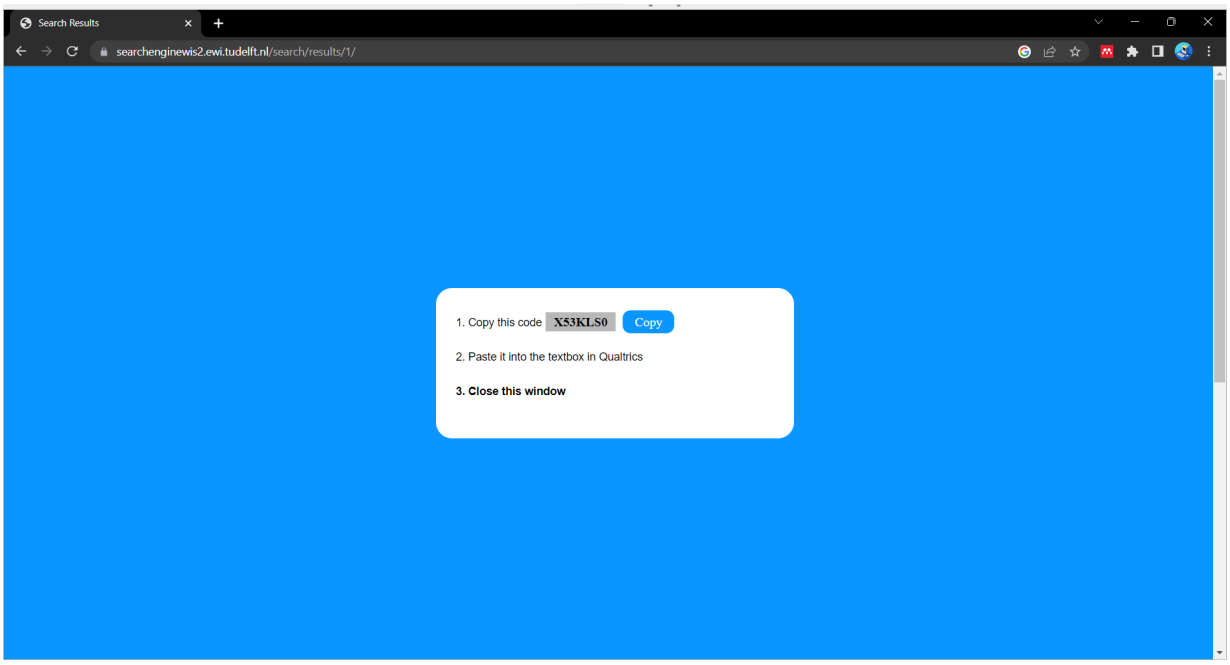


Figure B.4: Screenshot of the search engine. Upon returning to Qualtrics we showed the participant a completion code that served as an additional attention check.