Quantifying and Understanding the Effects of Cognitive Biases in Human-Computer Interaction

Nattapat Boonprakong
ORCID: 0000-0002-0735-4536

PhD Thesis

October 2025

School of Computing and Information Systems
Faculty of Engineering and Information Technology
The University of Melbourne

Submitted in total fulfilment of the requirements for the degree of Doctor of Philosophy.

This page is intentionally left blank.

Abstract

Modern technologies enable new and complex ways for humans to interact with computers. They tend to impose cognitive demands, time constraints, and ambiguity on users. To cope with such demands, humans apply mental shortcuts to sift through information and effectively make decisions. These shortcuts result in cognitive biases, a concept proposed by Tversky and Kahneman as systematic, automatic tendencies that influence our behaviour and judgment. These biases can both introduce harmful effects and offer swift mental strategies to form good decisions. When it comes to human-computer interaction (HCI), cognitive biases can influence how users engage with computing systems. To better understand this interplay, this thesis forms a systematic understanding of how cognitive biases manifest in HCI. Informed by a scoping review of HCI articles that study cognitive biases, we found that computing systems can be designed to trigger, mitigate, and capitalise on the effects of cognitive biases.

This thesis provides grounds for conducting HCI research on cognitive biases, in which we tackled three challenges: quantifying the effects of cognitive biases, understanding how cognitive biases manifest in the user-system interaction, and designing systems that take cognitive biases into account. In brief, we explored the potential of physiological measurements, especially hemodynamic activity, as an indicator of cognitive biases. We found that cognitive biases do not manifest in every individual and context. Subsequently, we proposed the notion of cognitive bias susceptibility to account for individual and contextual factors that amplify and mediate the effects of cognitive biases. We also note that these factors can be taken into account when designing interventions to mitigate harmful cognitive biases. Finally, we formulated the understanding of cognitive biases into a blueprint of computing systems that encompass bias-awareness: the ability to detect and address cognitive biases that surface in HCI. We describe affordances that allow users to interact with systems without engaging with mental shortcuts that lead to problematic biases, e.g., sharing misinformation or relying predominantly on ideological beliefs.

Our findings motivate the need to bridge the cognition gap between humans and computers. Computing systems, if not carefully designed, can put cognitive demands, such as information overload and time constraints, rendering a fertile environment for problematic cognitive biases. We can also design computing systems to intentionally trigger cognitive biases that benefit users, for example, to facilitate behaviour change. On the other hand, these biases can be abused against the good of people as they open doors for behavioural manipulation. We discuss societal and ethical considerations for designing bias-aware computing systems. We also emphasise that the HCI community should engage with the ongoing discussion in psychology and behavioural science with respect to the evolving definition of cognitive biases.

This page is intentionally left blank.

Declaration

This is to certify that:

1. this thesis comprises only my original work towards the degree of Doctor of Philosophy;
2. due acknowledgement has been made in the text to all other material used;
3. appropriate ethics procedure and guidelines have been followed to conduct this research;
4. the thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies, and appendices.

Nattapat Boonprakong
October 2025

This page is intentionally left blank.

Preface

This thesis is submitted in total fulfilment of the requirements for the degree of Doctor of Philosophy at the University of Melbourne. The entirety of the work presented in this thesis was conducted during my PhD candidature at the School of Computing and Information Systems, under the supervision of Associate Professor Tilman Dingler, Dr Benjamin Tag, and Associate Professor Jorge Goncalves.

This thesis comprises four peer-reviewed articles, of which I was the primary author. I proposed the research questions, planned the study designs, developed the software for data collection and analysis, recruited participants, and collected the user study data. I drafted the research articles for submission and revised them based on the feedback received from the peer-review process. I appreciate the contributions of the listed co-authors, who provided feedback to refine the study designs, data analysis methods, write the manuscripts, and address revisions. Therefore, I use the scientific term "we" throughout the main chapters of this thesis in recognition of my co-authors' contributions.

The research articles included in this thesis are referred to by Roman numerals (Articles I, II, III, and IV) as indicated below. All articles are included in full (author-approved version), preceded by a brief introduction situating each article within the context of this thesis.

Article I Nattapat Boonprakong, Benjamin Tag, Jorge Goncalves, and Tilman Dingler. 2025. How Do HCI Researchers Study Cognitive Biases? A Scoping Review. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3706598.3713450 (*Received the Honourable Mention Recognition*)

Article II Nattapat Boonprakong, Xiuge Chen, Catherine E. Davey, Benjamin Tag, and Tilman Dingler. 2023. Bias-Aware Systems: Exploring Indicators for the Occurrences of Cognitive Biases when Facing Different Opinions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3544548.3580917 (*Received the Honourable Mention Recognition*)

Article III Nattapat Boonprakong, Saumya Pareek, Benjamin Tag, Jorge Goncalves, and Tilman Dingler. 2025. Assessing Susceptibility Factors of Confirmation Bias in News Feed Reading. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3706598.3713873

Article IV Nattapat Boonprakong, Benjamin Tag, and Tilman Dingler. 2023. Designing Technologies to Support Critical Thinking in an Age of Misinformation. In *IEEE Pervasive Computing*, July-Sept 2023. IEEE, Vol. 22, No. 3, pp. 8-17. https://doi.org/10.1109/MPRV.2023.3275514

The studies presented in Articles II and III received the necessary ethical approval from the University of Melbourne Human Research Ethics Committee (Ethics Application ID: 1956072.1). No other third-party editorial assistance was provided in preparation for this thesis.

This page is intentionally left blank.

Acknowledgements

First of all, I would like to thank my supervisors—Tilman Dinger, Benjamin Tag, and Jorge Goncalves—for guiding me throughout four years of PhD research. For me, there is no better combination of supervisors. I could not express the amount of encouragement, patience, and support they have given me. Thank you for believing in me and being a significant part of my academic journey.

To Tilman, thank you for taking me in and training me *academic muscles* – the term you mentioned when I first started in 2021. Thank you for never giving up on me, and always encouraging me to fight, through the hard times – surviving the pandemic, writing my first paper, thriving through an existential crisis, and all the transitions I had. I appreciate the academic and life lessons you have given to me: time management, mentorship, professionalism, and positive thinking. I have learned to aim for the best, prepare for the worst. I will always remember our days, and especially the very first day when you met me on Zoom. Thank you for bearing with me through a Karaoke room interview!

To Ben, thank you for your continued mentorship and for taking care of my mental and social well-being. Your comment made me cry when I wrote and revised my first paper (sounds dramatic, right?). But it was a very good inoculation for my early academic career, where taking criticism is like breathing in the air. Thank you for giving me the strength to thrive in the academic world, and for being there for me whenever I need help and advice. I always enjoyed our chats and your German humour!

To Jorge, thank you for always being super supportive and for always looking for opportunities that might suit me. Thank you for showing me that doing research is not too difficult when we do it systematically. Thank you for making me feel supported in the group, for showing me the ropes to succeed in academia, and for helping me navigate through the very end of my PhD.

I am fortunate to have worked together with my collaborators. To Kaixin Ji, thank you for being my academic partner, who never made me feel isolated in Melbourne. Thank you for the countless hours we chatted and ranted. To Saumya Pareek, thank you for being my role model and always being on board for user studies and workshops. To Si Chen, thank you for your continued support all the way from the States. Special thanks to Catherine Davey for taking the time to refine my experimental design and for her kind support, especially access to the biomedical engineering lab. Thank you Mengmeng Wang for guiding me through fNIRS data processing, and Xiuge Chen for pioneering the grounds for my first user study. Special thanks to Gaole He, Danding Wang, Ujwal Gadiraju, Niels van Berkel, and Jiqun Liu for embarking on our CSCW '23 workshop. I also thank Ziyi Ye, Damiano Spina, Tuukka Ruotsalo, and Flora Salim for running a UbiComp '24 workshop together. Beyond my kind collaborators, I would like to thank the anonymous reviewers who provided feedback that led me to greatly improve my work. I have learned a lot from you *all*!

I am grateful for the support and company of everyone at the HCI group. I thank my Melbourne Connect GR friends Ying Ma, Siquan Zhang, Samangi Wadinambiarachchi, Jarod Govers, Shaona Cheng, Songyan Teng, Piumi Perera, Cherie Sew, Ulan Kelesbekov, Le Fang, Suwani Gunasekara, Yan Zhang, Jian Zhang, Sara Schoembs, Sara Khorasani, Tharaka Ratnayake, and many more. I also appreciate the support of the senior

(ex-)members of the lab: Joshua Newn, Weiwei Jiang, Jing Wei, Difeng Yu, Qiushi Zhou, Ebrahim Babaei, Mo Zhang, Kangning Yang, Zhanna Sarsenbayeva, Gabriele Marini, and Brandon Syiem. Thank you for giving me a warm welcome when I first arrived in Australia. Special thanks to Andrew Irlitti, Jarrod Knibbe, Eduardo Velloso, Wafa Johal, Vassilis Kostakos, Ryan Kelly, Audrey Balaska, Adélaïde Genay, Sarah Webber, Wally Smith, Frank Vetere, Jenny Waycott, Melissa Rogerson, Greg Wadley, Bin Chen, and Arzoo Atiq. I would also take this opportunity to thank our lab managers, Allen Pilares and Antony Chacon, for being very helpful and supportive of my user studies in the UX lab.

Special thanks to the people at the School of Computing and Information Systems (CIS) for making my PhD candidature a memorable experience. To Trina Dey and Thanh-Dat Nguyen, thank you for your amazing work as the legendary teaching team of COMP90041. It was a pleasure to have worked with both of you. To Abby Yuan, Archana Vadakattu, Philip Cervenjak, Stella Peng, and Peter Wang, thank you for giving me the awesome CIS-GReS committee experience. To Liuliu Chen and Harindu Ashan, thank you for the fun times in organising the CIS Doctoral Colloquium. I thank my PhD advisory committee chair, Lars Kulik, for taking the time to ensure my PhD journey is as smooth as possible. I would also like to thank the CIS admin team, including Melissa Hofsteter and Laura Juliff, for their kind and swift support.

Further, I acknowledge the University of Melbourne and the Faculty of Engineering and Information Technology for providing me with the Melbourne Research Scholarship. I would like to thank the students I have tutored and mentored over the years, as well as all participants in my user studies.

To people who laid the academic foundations prior to my PhD: I would like to thank Proadpran Punyabukkana for introducing me to Human-Computer Interaction back in my undergraduate. Thank you for showing me the human-centred aspect of computing, which subsequently led me to pursue a PhD in this field. I thank my master's thesis supervisors, Tsukasa Kimura and Masayuki Numao, for giving me opportunities to establish my skills in physiological computing. Thank you for always keeping me motivated in doing research.

I would like to thank the people who have made my life in Melbourne can't be more enjoyable. To Nixon Wong, thank you for keeping me company and helping me to significantly improve my Thai cooking skills. To Lin Song, thank you for always making me feel supported and valued. I enjoyed the intellectual conversations we had over food, coffee, and drinks. Special thanks to Timothy Tuan, who supplied me with the much-needed emotional support all the way from Malaysia.

A big thank you to my family in Melbourne and Bangkok. I thank Ratawan and Steven Upton—my sister and brother-in-law— who are always there for me. Thank you for making my life in Melbourne feel like I have never left home. I thank my brother for always believing in me. To Mum and Dad, thank you for your unconditional love, sacrifices, and support in pursuing what and where I want to be.

Contents

Chapter 1

Introduction

Humans are imperfect and subjective creatures. Our cognitive and memory capacity are inherently limited. On the other hand, the world we live in is eminently complex. We apply instincts, gut feelings, and rules of thumb to sift through the complexity of the world. In other words, these so-called **"mental shortcuts"** allow us to efficiently make decisions in the real world. However, these shortcuts have shortcomings as they restrict our ability to make objective decisions. Sometimes, following gut feelings results in mistakes and rules of thumb do not always work (i.e., giving an optimal decision). Hence, the way we think, react, and behave is subjective and individualistic. As humans, we derive mental shortcuts as part of our survival and natural selection [131]. We constantly develop mental shortcuts through our past experience of the world [70].

Prominent psychologist Herbert Simon coined the concept of **bounded rationality**, which explains that humans do not always make rational decisions [157]. Simon believed that processing information (and making decisions) is computationally expensive. Humans do not have sufficient cognitive bandwidth, memory capacity, time to think, and knowledge of the world. Therefore, we are prompted to use mental shortcuts to filter and simplify the information we come across. These shortcuts serve as strategies to help us cope with the external reality while influencing how we react and form judgments according to our subjective worldview. However, oftentimes, these shortcuts lead to the distortion of our judgment and rationality.

Behavioural scientists and Nobel laureates Amos Tversky and Daniel Kahneman extended the idea of bounded rationality. They discovered different ways our mental shortcuts distort human behaviours. Specifically, Tversky and Kahneman showed that humans do not always conform to the rules of logic and probability. Instead, humans exhibit **cognitive bias** – the concept Tversky and Kahneman coined as a systematic deviation (or error) from the norm of rational judgment. Their line of research [87, 91, 177, 178, 180] systematically documented such patterns of deviation, for example, anchoring bias makes people rely heavily on the first information presented to them [178]; availability bias prompts individuals to rely on information easily available to them [177]; and the framing effect causes different reactions to a piece of information depending on how it is presented [180]. As mental shortcuts are hardwired to the human mind, cognitive biases are largely automatic and happen without our awareness.

The notion of cognitive bias is widely adopted beyond psychology and behavioural science. Scholars employ cognitive biases as lenses to explain human behaviour on both micro and macro scales. Doctors misdiagnose patients because they have confirmation bias. Designers fail to expand their ideas because they overrely on their limited ideas (i.e., design fixation). Juries make unfair judgments because they are influenced by various biases like stereotype, affinity, or anchoring. In election cycles, voters can fall victim to a plethora of cognitive biases (e.g., authority bias, halo effect, or availability bias). On the other hand, politicians take advantage of these automatic tendencies to *engineer* their campaigns to tap into people's cognitive biases, optimising the election outcome.

Without exception, cognitive biases surface when humans interact with computing machines. As a result, the user's interaction with computing systems produces systematic effects. Cognitive biases specifically emerge as users follow their hardwired, automatic tendencies when making judgments with computing systems. These tendencies span classic examples of cognitive biases. When searching for information on search engines, we tend to click on content items that confirm the hypotheses in hand. When interacting with an AI agent, there is a tendency to over-rely on the suggestions given by AI. When facing fake news on the Internet, people tend to integrate the news into their beliefs and be resistant to debunking. The field of human-computer interaction (HCI) has emerged to study systematic effects – as part of complex human behaviour when they use computers – and design user interfaces that take into account such effects to optimise the user experience.

The success of ubiquitous technology has made computers increasingly pervasive and integrated in parts of our everyday lives. While computing systems act as a companion to help us achieve tasks and make decisions, the design of these systems may not fit our complex and constrained minds. As a result, computing systems can trigger undesired effects when interacting with users. For example, the black-box design of AI systems can trigger a user's mistrust in the AI prediction. The design of rapid content sharing on social media platforms (e.g., the retweet button) makes it more lubricative for users to propagate unverified information. The mismatch between the design of computing systems and the human mind presents cognitive challenges to the users, who navigate in the world under their scarce cognitive and memory capacity, limited time, and insufficient knowledge. Computing systems tend to impose designs that overwhelm users with information, introduce a sense of urgency to act, and provide incomplete information to users. Therefore, users resort to mental shortcuts to meet the cognitive constraints. While these shortcuts allow us to efficiently navigate on the user interfaces, they can produce cognitive biases and lead to undesirable consequences. Real-world examples suggest elements of computing systems, algorithms, and user interfaces can trigger and amplify cognitive biases in users. Recommendation algorithms and search engines [2, 8, 176] tend to cater content items predominantly to the users' preferences and, therefore, can trigger confirmation bias (the tendency to seek information that only aligns with one's own beliefs). AI systems can trigger undesired cognitive biases in ways designers do not anticipate [127], such as automation bias (tendency to favour suggestions from automated systems [162]), anchoring bias, or framing effect.

By taking cognitive biases in users into account, we can design computing systems that keep the systematic effects *under control*. On one hand, computing systems can be designed to minimise cognitive biases, prompt reflection, and mitigate undesired effects. The awareness of cognitive biases informs designers to avoid choices of user interface that potentially trigger cognitive biases. System feedback can be incorporated to help alleviate the user's cognitive burden. Interface designs can support users to shift away from mental shortcuts and reach informed decisions. On the other hand, computing systems can be deliberately engineered to take advantage of cognitive biases, steering the user's behaviour and decision-making. Online misinformation and viral content capitalise on cognitive biases to gain engagement and propagate on the Internet. Dark patterns offer the design of user interfaces that target users' cognitive biases to sway their behaviours in a predictable way, for example, to make a subscription or buy an item [116, 117]. Similarly, social engineering applications exploit cognitive biases to influence people's decision-making on a large scale [19], like the infamous 2016 Cambridge Analytica scandal [12].

Cognitive biases arise as a by-product of the interaction between humans and computers. They influence how users interact with computing systems and introduce concerns in the real world. It is, however, a challenge to build systems that address and give the user control over these biases. We neither have a sufficient and systematic understanding of how cognitive biases manifest in human-computer interaction, nor pursue a method to precisely detect these biases in the interaction.

1.1 Research Questions and Contribution

This thesis aims to enrich the systematic understanding of the effects of cognitive biases on HCI, develop methods to detect these effects, and sketch a blueprint for the design of computing systems that take into account cognitive biases that arise in human-computer interaction (HCI). This motivates the overarching research question in this thesis:

How do cognitive biases manifest in HCI?

Cognitive biases have a profound and pervasive impact in the real world. In the same vein, these biases are present in many scenarios of human-computer interaction. However, we lack a clear understanding of the role of computing systems, algorithms, and user interfaces in mediating the manifestation of cognitive biases. Specifically, HCI researchers borrow the notion of cognitive biases from behavioural science and psychology to explain effects in user-system interaction. While the issue of cognitive bias has increasingly been discussed in HCI, researchers employ different angles, methodologies, and application contexts to conduct cognitive bias research. Therefore, we lack a systematic understanding of how HCI researchers engage with these biases. We ask the first research question of this thesis:

RQ 1: How are cognitive biases studied in HCI?

We address **(RQ 1)** by conducting a scoping review to map out cognitive bias studies in HCI published between 2010 and 2024. The findings of this review provide a survey contribution to HCI: we identify the research conduct and gaps in the literature (Article I). Based on the analysis of our article corpus, we found evidence that computing systems can trigger cognitive biases in users and influence how users interact with computing systems. Designers may take advantage of cognitive biases in users and build user interfaces that leverage biases to steer user behaviour. HCI researchers develop tools and methods to closely observe this phenomenon.

Furthermore, we derive three main narratives where HCI researchers engage with cognitive biases, as shown in Figure 1.1. Motivated by real-world concerns about cognitive biases influencing human behaviours, HCI researchers develop tools and methods that capture the occurrences of cognitive biases to closely study their effects on the interaction between humans and computers. Subsequently, the understanding of cognitive biases informs the design of computing systems and user interfaces, which take cognitive biases into account, mitigate undesirable biases, leverage their useful effects, and, in turn, help address problematic behaviours of humans in the real world.

The findings of this scoping review also highlight one prominent research gap: the lack of tools and methods to quantify the effects of cognitive biases in HCI (Narrative 1 in Figure 1.1). With the ability to detect the occurrences of cognitive biases, designers can precisely capture moments where biases surface and prompt interventions to address these biases specifically. One main challenge is that cognitive biases largely happen without users' awareness [130, 199]. The subconscious nature of biases makes it difficult to obtain an objective ground truth for their occurrences. While measures of cognitive biases exist from psychology, for example, the Wason selection task [190] for identifying confirmation bias or the implicit association test [63] for measuring implicit bias, limited work explores the measurement of cognitive biases in user interaction with computing systems. Therefore, this thesis asks the second research question:

RQ 2: What are the indicators for the occurrences of cognitive biases when they manifest in HCI?

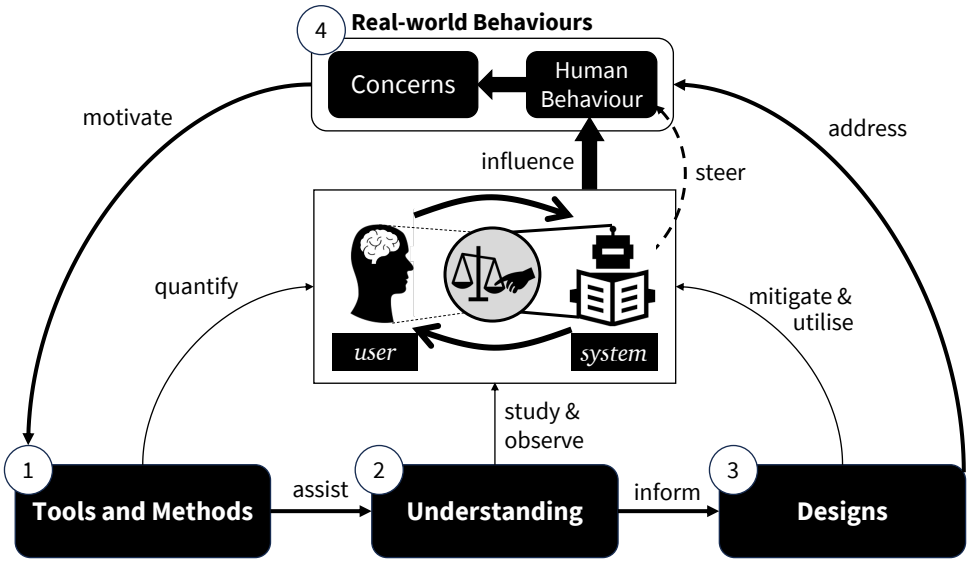


Figure 1.1: This diagram charts three narratives of how HCI researchers engage with cognitive biases: (1) developing tools and methods to quantify the effects of cognitive biases, (2) conducting empirical studies to understand and observe cognitive biases in human-computer interaction, and (3) deriving designs of computing systems that take into account cognitive biases in people and address real-world concerns arising from human behaviours.

To address **(RQ 2)**, we look into the common scenario of information consumption where users encounter expressions on a polarising, divisive issue. Ideologically polarising information can trigger cognitive biases as individuals rely on their existing beliefs to process such information. We designed and conducted a study where participants were exposed to messages containing opinions on a divisive, polarising topic. The exposure to these messages can trigger cognitive biases. At the same time, we explored measures for the occurrences of cognitive biases – namely, self-report, behavioural, and physiological measures. The findings of this study (Article II) provide an empirical contribution to HCI: we present a study design that triggers cognitive biases in opinion reading and show that physiological signals, in particular hemodynamic activity measured through Functional Near-Infrared Spectroscopy (fNIRS), indicate the effects of cognitive biases in such scenarios.

The findings of Article II interestingly suggest that the effects of cognitive biases are not always pronounced in every individual and setting. We found that cognitive biases only influenced hemodynamic activity in participants who exhibited low interest in the topic. This is supported by works in psychology, which suggest that cognitive biases do not occur independently: a plethora of user-related factors can either facilitate or hinder the occurrences of cognitive biases [138, 139]. Recent investigations on cognitive bias mitigation [21, 140, 143] report that a variety of factors related to user and interaction-context characteristics influence the effectiveness of bias mitigation interventions. These factors, therefore, determine users' susceptibility to exhibit cognitive biases. Limited research investigates the interplay between user- and context-related factors, the occurrences of cognitive biases in HCI, and the effectiveness of interventions to mitigate cognitive biases. By filling this research gap, we can enrich *the understanding* of cognitive biases (Narrative 2 in Figure 1.1) by considering user- and context-related factors. Therefore, this thesis asks the third research question:

RQ 3: What and how factors for user- and interaction-context influence the occurrences of cognitive biases when they manifest in HCI?

	Research Question	Addressed in	Narrative
RQ 1	How are cognitive biases studied in HCI?	Chapter 3	General Outline
RQ 2	What are the indicators for the occurrence of cognitive biases when they manifest in HCI?	Chapter 4	1. Tools & Methods
RQ 3	What and how do factors for user- and interaction-context influence the occurrences of cognitive biases when they manifest in HCI?	Chapter 5	2. Understanding
RQ 4	What are the considerations for the design of computing systems that take cognitive biases into account?	Chapter 6	3. Designs

Table 1.1: Research questions in this thesis, their corresponding chapters, and the narratives they address.

To address **(RQ 3)**, we conduct a user study that assesses the influence of individual and contextual factors on the occurrences of confirmation bias when individuals consume tweet-like information through a news feed. With a similar setting to the second study reported in Article II, we induced confirmation bias as individuals skim through information content on a polarising, divisive topic. The findings from this study (Article III) provide an empirical contribution to HCI: we found that an individual's thinking style, political beliefs, perception of the content's issue strength, and the task design influence the occurrences of confirmation bias. Further, we provide practical implications for designing interventions that take into account individual and contextual factors to effectively mitigate cognitive biases when consuming information online.

With a better understanding of how cognitive biases take effect in the interaction between users and systems, we can derive implications to inform the design of computing systems that address and take into account cognitive biases that exist in users and emerge from their interactions with systems. However, we do not have a clear research space on how to design these systems to effectively keep up with the dynamics of cognitive biases. This presents a research gap regarding the design of computing systems (Narrative 3 in Figure 1.1). Therefore, this thesis asks the fourth research question:

RQ 4: What are the considerations for the design of computing systems that take cognitive biases into account?

To address **(RQ 4)**, we gather insights from three academic workshops that brought together researchers from multiple disciplines to discuss the research space for the issue of cognitive biases in HCI with a special emphasis on the issue of online misinformation, which cognitive biases take a major role in promoting its spread. The findings provide an opinion contribution to HCI: we propose a research agenda to design technologies that equip users with affordances to make informed decisions and boost their skills to be resilient against manipulation when navigating the online world (Article IV). We also signal to the community to consider the issue of cognitive biases from interdisciplinary lenses, as well as conduct research that is grounded in theory.

1.2 Research Methodology

In this thesis, we refer to cognitive biases as a human factor that systematically influences user behaviour and decision-making. We specifically define these biases as a by-product of heuristics that users employ to

navigate the user interface. In psychology and behavioural science, the definition of cognitive bias has been subject to debate. Although Tversky and Kahneman [178] originally defined cognitive biases as systematic deviation from the norm of rational judgment, other schools of researchers refer to the notion of biases and heuristics differently, for example, Gigerenzer [52] viewed cognitive biases as rather useful strategies to make sense of the world than cognitive errors; and Haselton [70] believed that cognitive biases are the design features of the human mind. The studies presented in this thesis, therefore, approach cognitive biases as lenses to study systematic effects in human-computer interaction.

In Article I, we employed a systematic scoping review [123] as a method to chart the narratives of cognitive biases in HCI, aiming to address **(RQ 1)**. Because the research space of this issue was previously unexplored, the scoping review methodology aims to identify and map the body of the literature. We followed the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statement [122] to identify and screen articles based on our eligibility criteria (i.e., articles that study cognitive biases). Following prior HCI surveys (e.g., [23, 168]), we included publications from venues representative of HCI research (i.e., all ACM SIGCHI-sponsored venues). Furthermore, we performed open coding to classify articles in our corpus based on study focus and application context. One researcher iteratively performed the coding analysis to derive the study focuses and application contexts, with three other researchers cross-checking the process.

In Articles II and III, we operationalised cognitive biases in the context of online information consumption, where individuals often encounter information expressing opinions on a polarising, divisive topic. In this regard, we define heuristics as the user's reliance on the alignment between their existing beliefs and the content's ideological stances. These heuristics result in cognitive biases – systematic effects on the user's interaction with computing systems. We employed text opinions on different polarising topics as stimuli to induce the application of these heuristics and, thus, the effects of cognitive biases. Particularly, we derived these text opinions from Britannica's *ProCon.org*, which hosted opinion or factual quotes that either support or oppose an ideology. We selected a subset of topics that were appropriate for our research participants in Australia and domestically dominant in the public debate. We appropriated and simplified the text stimuli to fit in a tweet (50-80 words in length, written in English). In the studies, we did not disclose the source and author of the tweets as well as their ideological stances to minimise confounding effects, for example, source and authority bias. We presented the stimuli on a screen and deployed the study flow on progressive web applications, e.g., *Qualtrics*[1]. We ensured that the presentation of the text stimuli was comfortable to read and did not introduce a significant mental burden to the participants. We recruited the participants from the university community, with a requirement that they were over 18 years old and spoke English as their first language. We used a Likert-scale questionnaire to gauge the user's ideological beliefs and topic interest in different topics employed in our studies. In Article III, we also employed the word association test [103, 110] as an alternative measurement of the user's ideological beliefs. We cross-checked between two different measures and found that they are highly correlated (Pearson $\rho = 0.946$).

In Article II, we used self-report, behavioural, and physiological measures to detect the occurrences of cognitive biases, aiming to address **(RQ 2)**. We conducted two user studies, with the design of Study 2 building on Study 1 to minimise confounds and follow up on its inconclusive findings. In Study 2, we collected the user's agreement with the ideological stance of the tweet (Q1), the user's likelihood to share the tweet on their social media (Q2), and the user's effort spent reading the tweet (Q3). All self-report measures were gathered as Likert-scale data. In our analysis, we cross-checked the self-report measures with behavioural and physiological measures to ensure the internal validity of our findings. We collected eye-tracking data, electrodermal activity (EDA), and hemodynamic signals (fNIRS). Therefore, both studies presented in this article were conducted in a controlled lab environment to allow appropriate collection of behavioural and physiological measurements. We followed appropriate practices and pretreatments for EDA data collection and handling in HCI studies [9]. Physiological responses generally have a few-second delay after an exposure

[1] www.qualtrics.com

event to a stimulus. Therefore, we provided a minimum 15-second break as an inter-stimulus interval to allow a sufficient gap between two stimulus exposures, avoiding the physiological responses from confounding each other. Prior to data collection, we calibrated the sensing devices and allowed time (5-10 minutes) for EDA and fNIRS signals to reach a baseline. We applied data pre-processing and cleansing to remove motion artefacts that occurred during the studies. For data analysis, we applied one-way repeated measures ANOVA to examine the effects of user-tweet ideological congruency on all measures. We specifically considered the participant ID as a random effect in the analysis to address individual differences.

In Article III, we explored individual and contextual predictors that influence the effects of confirmation bias. Measured through self-report, Likert-scale measures, these predictors are candidates for factors in the user's susceptibility to cognitive biases. We deployed the text stimuli in three information consumption scenarios: information seeking, recall, and interpretation. Addressing **(RQ 3)**, we examined the interaction effects between the user-tweet ideological congruence (i.e., confirmation bias) and the bias susceptibility predictors. We regard significant interaction effects as evidence for the predictor being a factor in cognitive bias susceptibility. We derived the measure of confirmation bias in each scenario from the prior literature [100, 183]. In particular, we employed the headline ranking task, free-recall task, and information interpretation questionnaire as measures for confirmation bias in information seeking, recall, and interpretation scenarios. For the free-recall task, we assessed the free-recall ability using a rating approach. Two researchers individually gave a score for the completeness of the participant's recall response on a Likert scale. We ensured the consistency of the ratings by using the inter-rater reliability (Cohen's Kappa). We conducted the study as an in-lab experiment to control the reliability of our data by making sure our participants followed our procedures and were not distracted. We also performed a priori power analysis to determine the minimum required sample size with a medium-to-large effect size. For data analysis, we performed mixed-effects ordinal regression using Cumulative Link Mixed Models (CLMM) to assess the interaction effects, while accommodating the ordinal-scale data and considering the participant and stimulus ID as random effects.

In Article IV, we applied open coding to extract research themes from discussions in three workshops, addressing **(RQ 4)**. We first analysed each workshop's content, which we gathered from two sources: *Miro*[2] virtual boards used in the workshops and participants' position papers from each workshop's online archive. In each workshop, participants engaged in *action groups*, where they received challenges and solved them as a group of 4 to 6 individuals. We aligned the workshop notes with the action groups and challenges they aimed to address. Using open coding, we derived notes into research challenges and proposals for each workshop. Similar to our approach in Article I, one researcher iteratively performed the coding analysis, and two other researchers cross-checked the process.

1.3 Ethical Considerations

We took several steps to ensure the appropriate ethical conduct of the research in this thesis. The experimental design of the studies presented in Articles II and III received ethical clearance from the University of Melbourne Human Research Ethics Committee. Our research participants were informed in advance about the purpose and objective of the study. We provided them ample time to read a Plain Language Statement that explained each individual study's objectives, the data that would be collected, how the collected data would be stored and how it would be used for research, and potential benefits of the study to the participant, the society, and the research community. Participation in our studies was completely voluntary. Participants could withdraw from the study at any time without consequences. We obtained the participants' consent before starting the experiment. We compensated all participants for their time at the current minimum hourly wage in Australia.

[2] www.miro.com

Stimuli we deployed in Articles II and III contained expressions on a divisive topic and, therefore, may trigger stress and emotional responses. While aiming to select topics that are dominant and relevant to everyday debates in Australia, we avoided using overly sensitive topics that might bring discomfort to our participants. We informed the participants about the topics they would encounter in our studies, and, upon completion of the data collection, we debriefed them individually to clarify the real objective of the experiment. We explained to them that the stimuli were to trigger their cognitive biases, which we observed in our studies. We asked all participants if they wished to remove their data from the experiment, but none of them did so. The collected data were anonymised prior to the analysis. We did not collect or store identifiable information of the participants, apart from their demographics relevant to the study objectives (e.g., age, gender, political concordance, and physiological responses). We stored the collected data on a firewall-protected, secure server that can only be accessed by authorised users.

We collected physiological data as part of the studies in Article II. We informed our participants in advance that they would need to wear physiological sensing devices and ensured that they experienced minimal discomfort during data collection, which required the EDA electrodes and fNIRS optodes to be attached to the participant's skin. We opted for wearable, non-intrusive devices that can be comfortably attached like body accessories – i.e., wristbands and forehead patches. For some participants, we explicitly asked them to adjust their hair to remove hair artefacts, which can confound fNIRS data. We acknowledge that the experimental setting required our participants to refrain from moving their bodies to minimise motion artefacts. To minimise fatigue caused by the prolonged study session and the presence of the sensing devices, we provided a minimum 15-second break between two stimulus exposures. The breaks allowed our participants to rest and refresh (e.g., they could move their bodies, drink water, or chat with the experimenter). We discarded data collected during the break period.

1.4 Thesis Outline

The rest of this thesis is organised as follows. Chapter 2 introduces theoretical concepts that ground the notion of cognitive bias and discusses prior research that investigates the impact of cognitive biases in HCI.

Chapter 3, 4, 5, and 6 present four research articles that address the research questions proposed in this thesis. Chapter 3 describes Article I, which presents a scoping review of 127 HCI articles that study cognitive biases. Addressing **(RQ 1)**, we chart how HCI researchers engage with cognitive biases, document the research conduct, and derive narratives for the manifestation of cognitive biases in HCI (Figure 1.1). Chapter 4 features Article II, which explores different tools and methods to quantify the in-situ effects of cognitive biases in the context of information consumption. We address **(RQ 2)** by examining the reliability of self-report, behavioural, and physiological measurement as indicators for the occurrences of cognitive biases. Article II presents two user studies where users were exposed to ideologically polarising opinions, potentially triggering their cognitive biases, while we monitored their behavioural and physiological responses. Chapter 5 presents Article III, which investigates factors for user and interaction context that influence the effects of cognitive biases in the context of information consumption. We address **(RQ 3)** through a user study that examines the influence of user and context predictors for bias susceptibility on the effects of confirmation bias in three information consumption scenarios: information seeking, recall, and interpretation. Chapter 6 presents Article IV, which summarises scholarly discussions from three workshops around the issue of cognitive biases in HCI with a special focus on misinformation. We address **(RQ 4)** by looking into insights from the workshops and forming a design space of computing systems that address cognitive biases in their users.

Chapter 7 then combines the findings from different perspectives, describes how they answer the research questions of this thesis (**RQ 1, RQ 2, RQ 3,** and **RQ 4**) and reflects on how they contribute to HCI research. Furthermore, it proposes several avenues for future research, informed by the findings in Articles I–IV.

Chapter 2

Background

In this chapter, we review the literature on the issue of cognitive biases in HCI that informed the research questions of this thesis. We introduce concepts in psychology and behavioural science that are vital to understanding the notion of cognitive bias in Section 2.1. In section 2.2, we review the HCI literature that grounds the notion of cognitive bias.

2.1 Related Concepts of Cognitive Bias

Cognitive biases refer to deviations from the norm of rational judgment. These biases stem from natural constraints in decision-making. First of all, humans face uncertainty as they do not have complete information about the real world. Second, humans pursue limited cognitive capacity and time to make decisions. Therefore, humans resort to using heuristics – simple rules of thumb to simplify the problem – to make *quick, good-enough* decisions. While heuristics are useful cognitive strategies, they systematically *bias* humans' behaviours without their awareness. Since the seminal work of Tversky and Kahneman [178], researchers in psychology and behavioural science have extensively employed the concept of cognitive biases to explain the dynamics of human cognition and behaviours. In this section, we review the foundational ideas of cognitive bias: we discuss the concepts of bounded rationality, heuristics, and the dual process theory. More specifically, we discuss the definition of cognitive biases, which provides a theoretical ground for this thesis.

2.1.1 Bounded Rationality

Aristotle defined rationality as the crucial, unique characteristic of humans, differentiating from animals and plants [98]. More specifically, humans pursue the ability to use reason and logic to make decisions. Traditional economics make a grand assumption of *homo economicus* (in Latin for "economic man") that humans are perfectly rational and constantly make optimal decisions based on their complete knowledge of the world [173]. Philosopher John Stuart Mill [121] described a rational being as a self-interested agent who seeks to maximise their personal utility. Pioneers of game theory, John von Neumann and Oskar Morgenstern [186], developed Mill's idea into the theory of *expected utility* to formalise how humans make decisions. When considering the fact that humans often face uncertainty and have incomplete information about the world, humans make decisions based on probability and maximise the expected utility of the choices.

Critiques of Perfectly Rational Humans

Empirical research in behavioural economics and psychology challenged the homo economicus assumption and the expected utility theory, showing that human judgement can deviate from rationality and the rules of probability. For example, economist Maurice Allais suggested that humans do not always make decisions based on the expected utility [3]. Allais, in his work, presented an experiment (later famously known as the Allais paradox) which asks individuals to select between two choices (A or B) in two gambles (Gamble 1 and 2):

- **Gamble 1, choice A**: A 100% chance of receiving $1 million.
- **Gamble 1, choice B**: A 10% chance of receiving $5 million, an 89% chance of receiving $1 million, and a 1% chance of receiving nothing.
- **Gamble 2, choice A**: An 11% chance of receiving $1 million, and an 89% chance of receiving nothing.
- **Gamble 2, choice B**: A 10% chance of receiving $5 million, and a 90% chance of receiving nothing.

Allais showed that most individuals who took the survey preferred Gamble 1A and Gamble 2B. However, this particular combination does not give the maximum expected payout. Instead, choosing Gamble 1B and 2B will gain the maximum expected payout[1]. Amos Tversky and Daniel Kahneman [91] further extended Allais's finding and conducted an experiment where 72 individuals were asked to perform two tasks in the following:

- First, participants considered **Problem 1**: they chose between:
  A. Gain 2,500 with probability of 0.33, gain 2,400 with probability of 0.66, and gain nothing with probability of 0.01
  B. Gain 2,400 with a probability of 1.00
- Subsequently, participants were presented with **Problem 2**: they chose between:
  C. Gain 2,500 with a probability of 0.33 and gain nothing with a probability of 0.67
  D. Gain 2,400 with a probability of 0.34 and gain nothing with a probability of 0.66

Tversky and Kahneman observed that 82% of the participants chose Option B in Problem 1 and 83% chose Option C in Problem 2. Not only did the findings not conform to the expected utility theory, but they also showed that humans do not always make decisions based on the rules of probability. More specifically, their observations suggested that human decision-makers tend to define a reference point and frame the consequences, gains, and losses of the options based on deviations from their reference point. Problem 1 showed that humans can be *risk-averse* because the majority of participants selected Option B, which has a smaller expected gain ($\mu_A = 2500 \times 0.33 + 2400 \times 0.66 = 2409, \mu_B = 2400$). Tversky and Kahneman repeated similar observations and proposed the prospect theory [91, 181], which laid the foundation for cognitive biases.

[1] According to the expected utility theory, in millions$, the expected payouts for Gamble 1A+2A is 1.11, Gamble 1B+1B is 1.89, Gamble 1A+2B is 1.50, and Gamble 1B+2A is 1.50.

Simon's Bounded Rationality

Social scientist Herbert Simon [156] suggested that decision-making in the real world is computationally expensive. When making decisions, humans face constraints being their cognitive limitations (i.e., lack of knowledge and limited ability to forecast the future) and the complexity of the environment (i.e., difficulty of the problem and time limitations) [159]. Simon proposed the notion of *bounded rationality* [157], which states that humans would rather make a decision that satisfies the decision-making constraints than maximise the utility (i.e., finding the optimal solution). Simon and Newell further hypothesised that, because the cost of considering all possible solution alternatives is expensive, humans employ mental shortcuts or *heuristics* [161] to simplify the complexity of the problem and effectively make decisions that are good enough. In his seminal work [156, 157], Simon defined heuristics as "satisficing" – a combination of the words "satisfy" and "suffice" – to describe cognitive strategies humans make to form the best decisions that satisfy cognitive and ecological constraints they face in the real world. Satisficing heuristics, therefore, allows humans to make computationally efficient decisions under such limitations.

Simon's concept of bounded rationality calls for an alternative basis for modelling human behaviour and decision-making [158]. Compared to the expected utility theory, the concept of bounded rationality recognises the cost of gathering and processing information, and humans do not always meet such demands, and, therefore, make decisions that are not strictly optimal. Simon once described human behaviour as shaped by two-blade scissors: one being the human's cognitive capabilities and another being the structure of the decision-task environment [160].

Wason Selection Task

Psychologist Peter Wason devised a famous logic puzzle, the "four-card problem" (also known as the Wason selection task). One example of the puzzle states that:

You are shown a set of four cards placed on a table, each of which has a number on one side and a colour on the other. The visible faces of the cards show 3, 8, blue and red. Which card(s) must you turn over in order to test a rule that "if a card shows an even number on one face, then its opposite face is blue"?

In his work [192], Wason showed that less than 10% of study participants, however, found the correct solution (turning over the 8 and red cards). The solution can be deduced by classical logic (i.e., using modus ponens and modus tollens), which is to choose the cards to disconfirm the rule *IF a card shows an even number on one face, THEN its opposite face is blue.* Yet, Wason found poor performance on this selection puzzle. He hypothesised that individuals tend not to conform to logical thinking but instead avoid falsifying hypotheses they have in hand. Wason coined *confirmation bias* to explain such tendencies to ignore information that falsifies one's beliefs and hypotheses in hand [45, 191]. In the same line, other psychologists (for example, Fredrick in his well-known "bat and ball problem[2]" [49]) have shown that humans often fall victim to intuition when solving puzzles. Wason's puzzle not only supported the claim that human rationality is limited but also laid the foundation of the dual process theory and cognitive biases, which we review in the following.

2.1.2 Dual Process Theory

Psychologists widely recognise humans have two modes of thinking: *fast, intuitive* thinking and *slow, deliberate* thinking. The idea dates back to the 19th century in William James' book "The Principles of Psychology" [80], which suggested that humans had two distinct thinking processes: associative and true reasoning.

[2]The puzzle asks "a bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost?" Frederick found that half of the subjects gave the wrong, intuitive answer, 10 cents. (The correct answer is 5 cents)

James believed that associative knowledge was formed using past experiences, while true reasoning was used in new, unfamiliar scenarios.

Wason's selection task showed that most individuals relied on intuition rather than logic to make judgments. Building on their previous findings, therefore, Wason and Evans [190] proposed *Dual Process Theory*, suggesting that individuals possess two distinct thinking processes: heuristic process and analytic process. Stanovich and West [166, 167] labelled these processes as System 1 and System 2, and Kahneman [86, 87] called them intuition and reasoning. Modern psychologists have agreed that the former process is fast, automatic, effortless, associative, emotionally charged, and often difficult to control. In contrast, the latter process is slower, serial, effortful, and deliberately controlled [44, 46]. Kahneman [87] postulated that because humans have limited cognitive capacity, System 1 thinking tended to dominate judgment and reasoning, making them rely on previous experiences of the world. Meanwhile, human judgments were always explicit and intentional and, therefore, involved System 2 thinking to make novel judgments. When engaging in a demanding mental activity (e.g., attempting to hold in mind several digits), Kahneman suggested that individuals tended to shift from System 2 to System 1 thinking to offload their mental demands.

Another well-known version of the dual process theory is called the elaboration likelihood model of persuasion (ELM). In their theory, Petty and Cacioppo [20] proposed that there are two distinct routes people follow when they process information they come across: a central route and a peripheral route. The central route involves careful consideration and scrutiny of the message, while the peripheral route prompts individuals to rely on superficial cues, such as the credibility of the source or the attitude of the message, rather than the content of the message itself. The central route occurs when one's motivation and cognitive ability are high. On the other hand, individuals take the peripheral route when they do not have enough motivation or ability to engage with the message. Psychologists employ ELM to explain not only how individuals are persuaded by messages but also how they approach the information they receive to make decisions.

2.1.3 Heuristics and Cognitive Biases

Simon's idea of heuristics inspired behavioural scientists and psychologists to better explain bounded rationality. Tversky and Kahneman conducted a series of empirical experiments [86] and observed many heuristics humans employ to make decisions under uncertainty and computational intractability. Specifically, they coined the term *cognitive bias* to describe the phenomenon where heuristics lead to systematic error in judgment. In contrast, psychologist Gerd Gigerenzer [52] has a different view of heuristics as cognitive strategies to efficiently make decisions that adapt to real-world constraints. In this section, we review the discourse around the definitions of cognitive biases based on two competing schools of thought: Tversky and Kahneman's *Heuristics and Biases* and Gigerenzer's *Fast and Frugal Heuristics*. We also discuss modern definitions of cognitive biases as adaptive features of the human mind from the lens of evolutionary psychology.

Cognitive Biases as Systematic Errors

Extending the concept of bounded rationality, Amos Tversky and Daniel Kahneman[3] conducted a series of experiments and discovered different ways human judgments can deviate from the rules of probability [89, 177, 178, 180]. In their seminal paper "Judgement under Uncertainty: Heuristics and Biases," Tversky and Kahneman [181] reported evidence that humans exhibited error in probability assessment, suggesting three heuristics pertaining humans' deviation from the rules of probability: *representativeness* (tendency to rely on stereotype when judging the likelihood), *availability* (tendency to rely on immediate examples that

[3]Both received the 2002 Nobel Prize in Economics for their joint work on cognitive biases and prospect theory, which bridged the fields of economics and psychology and established the field of behavioural economics.

come into one's mind), and *adjustment and anchoring* (tendency to rely on the first piece of information). More specifically, they found that these heuristics lead to systematic error in judgment (in other words, deviation from the norm of rationality). They coined the term *cognitive biases* to explain such systematic errors. Cognitive biases can take a form depending on the heuristics humans rely on. Tversky and Kahneman's research program on heuristics and biases and the prospect theory [91] subsequently revealed more variations of cognitive biases, such as *loss aversion* (tendency to experience losses more severely than gains) and *the framing effect* (tendency to react differently to the same information depending on how it is presented).

Cognitive biases happen in everyday decision-making without one's awareness. Kahneman [86] endorsed the dual process theory and suggested that cognitive biases occur because individuals rely on intuition (System 1 thinking) and employ heuristics to make quick decisions, sift through the world's complexity, overcome uncertainty, and meet time constraints. Therefore, individuals generally are not aware of their own biases. Kahneman believed that cognitive biases are influential and pervasive in everyday decision-making because System 1 thinking dominates most human cognitive processes [87, 92]. More importantly, Tversky and Kahneman's notion of cognitive biases and the dual system theory not only influences research not only in the field of behavioural economics and psychology but also in other disciplines investigating human cognition, behaviour, and decision-making, for example, finance [155], management science [26], political science [5], law [194], medicine [151], and decision-support systems [7].

The impact of cognitive biases is profound and pervasive. Many social institutions require individuals to make decisions. Cognitive biases systematically skew a jury's judgment, the doctor's diagnosis, people's opinion-making, or a politician's decision, leading to devastating consequences in the real world. Subsequent research later discovered variants of cognitive biases. Extending beyond the initial list of cognitive biases introduced in Tversky and Kahneman's research program, scholars labelled existing (and newly discovered) psychological phenomena that happened in the real world as deviations from the norm of rationality. For example, *confirmation bias* [124, 191] (tendency to seek, interpret, and recall predominantly information that confirms one's beliefs), the *fundamental attribution error* [146] ((tendency to overattribute the behaviour of others based on their characteristics), the *Dunning-Kruger effect* [104] (tendency to overestimate one's ability despite lacking competence), or the *IKEA effect* [126] (tendency to disproportionately place a high value on products they partially created). Some cognitive biases, however, are minor variants of other well-known effects. For example, the IKEA effect is derived from the well-known *endowment effect* [172] (tendency to view the value of a good higher when viewed as a potential loss than when viewed as a potential gain); the *recency effect* [32] (the tendency to more easily remember what happened recently)) is a sub-form of the *peak-end rule* [93] (tendency to judge an experience based on how they felt at its most intense moments). To date, there are over 180 documented forms of cognitive biases [11].

Due to their diversity, some scholars have attempted to develop a taxonomy or classification of cognitive biases. The seminal work of Tversky and Kahneman [178] suggested three classes of heuristics: *representativeness*, *availability*, and *anchoring and adjustment*. These categories are, however, not mutually exclusive, as some cognitive biases, like the framing effect, can span all three classes of heuristics [6]. Carter et al. [26] derived a mutually exclusive taxonomy of nine biases in decision-making.

Cognitive Biases as Fast and Frugal Heuristics

Psychologist Gerd Gigerenzer has been critical of Tversky and Kahneman's idea of cognitive biases [53, 196]. While Tversky and Kahneman believed heuristics lead to flaws in decision-making, Gigerenzer rejects their idea, viewing heuristics in a way that humans employ them to achieve accurate judgments [55, 58]. Humans can apply heuristics deliberately, as opposed to the idea of subconscious heuristics endorsed by Tversky and Kahneman's school. Moreover, these heuristics are part of humans' natural adaptation to the complexity, limitations, and social norms of the environment. Gigerenzer established the *fast and frugal heuristics* research

program. In his seminal work [52], he reported a number of cognitive strategies that allow humans to make computationally efficient choices in real environments. For example, *immitation* (tendency to imitate the successful members of their communities [73]), *default rule* (tendency to follow the default rule when there is no apparent reason to do otherwise [175]) or $\frac{1}{N}$ *rule* (tendency to invest resources equally across all options regardless of each option's potential loss and gain [74]).

Gigerenzer disagreed with the view that humans are irrational and biased. He believed that it is inappropriate to characterise a deviation from the norm of rationality as a bias or an error. While human judgments do not conform to the rules of frequentist statistics (via examples shown in Tversky and Kahneman's work [90]), humans make use of Bayesian statistics to make accurate probability assessments [187]. Moreover, the wording and framing of the problem can lead to misinterpretation of the problem, being deemed as cognitive biases. For example, in Tversky and Kahneman's famous Linda problem [179]:

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. Which one of the following is more probable?

(a) *Linda is a bank teller.*

(b) *Linda is a bank teller and is an active feminist.*

In their experiment, Tversky and Kahneman showed that most individuals selected option (b). They pointed out that it was attributed to the *representativeness heuristic* where people focused on the descriptive details about Linda with a stereotype of a feminist while ignoring the broader statistical likelihood of her being a bank teller. Moreover, Tversky and Kahneman suggested that the *conjunction fallacy* caused people to consider the probability of a combination of events higher than the probability of a single event (i.e., option (a)). In response, Gigerenzer [4, 120], argued that the framing of the problem is misleading. The word "probable" can be interpreted in many ways, such as "what happens frequently" or "what is more plausible." He contended that the Linda problem is indeed a single-trial, subjective probability problem, and, thus, it is not appropriate to apply statistical norms (e.g., the conjunction rule for probability) to single-case judgments. On the other hand, by using Bayesian inference, the probability for option (b), as conditional probability, i.e., $P(A \mid B)$, is higher than conjunction probability, i.e., $P(A \text{ and } B)$ [39][4]. Gigerenzer and colleagues [56] suggested that Bayesian statistics are computationally less expensive than frequentist statistics, and, therefore, humans employ Bayesian statistics to make reasonable inferences. While human judgments may not conform to rules of probability, Gigerenzer argued that they may conform to other norms of rational judgments.

Gigerenzer also believed that simple decision-making heuristics could work as well as or better than complex and exhaustive algorithms [55]. For example, the *take-the-best* heuristic allows individuals to quickly decide between two alternatives based on the first cue that discriminates against them. By searching for the optimal solution using a decision tree, the take-the-best heuristic allows individuals to make an effective decision without knowing about all possible alternatives [57]. The *less-is-more* heuristic leads individuals to favour a lesser option when it is presented separately, but the better option when both options are presented together. In an experiment, Hsee [78] showed that individuals were more willing to pay for 7 oz scoops of ice cream in 5 oz cups (i.e., small ice cream in a smaller cup) rather than 8 oz scoops of ice cream in 10 oz cups (i.e., bigger ice cream in a big cup). When presented with an isolated option, individuals tended to favour the former option, believing they were getting a good deal. However, when presented with both options, individuals no longer preferred the lesser option. Gigerenzer and Brighton [54] argued that knowing more information or investing in more computation can decrease accuracy; therefore, human cognition relies on simple heuristics (e.g., the less-is-more effect) to be more accurate than strategies that use more information and time [144].

[4] A simple proof using Bayesian inference is available at https://allendowney.github.io/ThinkBayes2/chap00.html

Bias-Varience Trade-off

There is ample evidence of cognitive biases and judgment errors in humans. Deriving from statistics and machine learning, behavioural and cognitive scientists believe that there is a trade-off between bias and variance. In real-world decision-making, individuals make errors because their cognition is biased, their knowledge is limited (i.e., there is variance in data), and there is random noise in the decision-making process:

$$\text{Error} = \text{Bias}^2 + \text{Varience} + \text{Noise}$$

In their work, Gigerenzer and Brighton [54] suggested that human cognitive systems face this bias-variance trade-off when making any inference about the world. However, humans generally do not have complete information about the world. Instead, they generalise very well based on a small number of observations, and, therefore, keep the variance term under reasonable limits. Unlike machines, humans can learn and generalise from a few instances; for example, teenagers can learn how to drive without hitting a pole 10000 times [196]. While cognitive biases emerge as compensation for low variance in decision-making, human cognitive systems usually perform well and accurately in the real world. Thus, he believed that humans apply heuristics as strategies to minimise errors in decision-making. On the other hand, a more comprehensive knowledge of the world would mean that human cognition has high variance and less bias. Based on the bias-variance trade-off, it runs the risk of overfitting the past observations when making future predictions. Psychologists argued that cognitive heuristics and biased algorithms allow humans to make accurate decisions and make sense of the world despite incomplete knowledge [38, 71, 131].

In a 2021 book, "Noise: A Flaw in Human Judgment," Kahneman, Sibony, and Sunstein [94] commented that noise contributes to judgment error, in addition to cognitive biases. They termed noise as non-explanable inconsistency or variability in judgments, while cognitive biases are systematic errors. Unlike deterministic computer algorithms, humans exhibit noise when making decisions. For example, two doctors can give different diagnoses about the same patient. In their book, the authors suggested that noise increases in attitude when individuals make decisions collectively by the added difference between individual judgments.

Cognitive Biases as Adaptive Features of the Human Mind

Evolutionary psychologists Martie Haselton and colleagues [69, 70] viewed cognitive biases as the design features of the human mind. Humans adopt heuristics and exhibit biases as part of their evolution and natural selection. Hasselton and colleagues [70] classified cognitive biases into three classes: *heuristics*, *error management biases*, and *artefacts*. First, cognitive biases arise from the use of heuristics humans employ under natural constraints for information processing. Humans specifically develop heuristics as useful strategies to make sense of the world, however, they can have unintended effects on judgments. For example, they apply stereotypes and overemphasise personality over objective truth in decision-making [148]. Second, humans adopt error management biases from natural cognitive mechanisms that cause the least costly error. They refer to the error management theory (EMT) [68], which states that any cognitive mechanism can cause two types of error: a false positive (type I error: adopting a belief that is not true) or a false negative (type II error: failing to adopt a belief that is true). EMT predicts that humans adopt decisions that minimise the net effect of the error and effort, as part of their survival and evolution. Error management biases arise as these natural errors are inevitable when humans make decisions in the real world. Third, cognitive biases emerge from an inappropriate assumption of normative standards. While behavioural scientists showed that human judgments did not follow the rules of frequentist probability, human brains employ Bayesian probabilities, which are computationally simpler [76]. Natural human behaviours are, therefore, considered cognitive biases because frequentist statistics are treated as a normative standard of probability assessment [51]. Moreover, the

framing of the problem content affects the way humans approach decision-making and, thus, can mislead their choices. While Wason showed that humans exhibited poor performance when solving the four-card problem [192], subsequent research showed that the presentation and framing of the problem influenced the subjects' performance on the task [31].

Through the lenses of evolutionary theory, psychologist Lionel Page argued that many behaviours previously labelled as cognitive biases instead lead to optimal decision-making outcomes. In his recent book "Optimally Rational: The Good Reasons We Behave the Way We Do" Page [131] suggested that these biases are adaptive strategies that humans develop over a long process of evolution. If humans were actually irrational and dumb, we would have already gone into extinction. Instead, humans adopt cognitive strategies that are optimal for survival. Heuristics, dual-process decision-making, and cognitive biases are shaped by natural selection to perform optimally under natural decision-making constraints (e.g., perceptual limitations, costs of gathering and storing information, and the structure of the environment). Page described that heuristics and biases are a major feature, not a bug of the human mind. However, there is a substantial mismatch between the environment in the modern world and the world in which humans traditionally evolved [30]. Kaplan [95] suggested that features of evolved human behaviour no longer fit with the modern world. Page [131] elaborated in his book through an example of sweet and fatty foods: while they were useful in the ancient age when energy-rich food was scarce, they led to issues in the modern environment. In the end note, Page criticised that the field of behavioural science should move beyond documenting specific cognitive biases (i.e., deviations from a norm of rationality) in a range of different settings and shift towards searching for unifying principles behind the observed phenomena.

Summary

Ongoing debates over the definition of cognitive biases, heuristics, and human rationality have developed a nuanced understanding of how human cognition and behaviour work in the real world. Simon proposed a revolutionary concept of bounded rationality and heuristics to help humans make decisions under their mental and environmental constraints. Two schools of psychologists later developed their definitions of heuristics. Tversky and Kahneman believed that heuristics lead to systematic errors and deviations from the norm of rationality. They called such distortions cognitive biases and documented several forms of how human decision-making can deviate from the norm. Cognitive biases systematically influence behaviour and allow individuals to form their subjective reality. Scholars from different disciplines have employed Tversky and Kahneman's notion of cognitive biases to explain human behaviour and decision-making. On the contrary, Gigerenzer viewed heuristics as strategies for making fast and accurate judgments. He believed that it is inappropriate to consider humans irrational by asserting a norm of rational thinking. Cognitive biases do not represent errors in judgment but fast and frugal strategies to help compensate for the bias-variance trade-off, thus allowing humans to generalise and make accurate decisions under their limited knowledge of the world. Evolutionary psychologists argue that individuals employ heuristics and biases in everyday decision-making as part of human evolution and natural selection. Therefore, cognitive biases are adaptive features of the human mind to help us make sense of the world. Nonetheless, these evolutionary cognitive biases present concerns in the modern world because they may no longer fit with the present environment with different incentives. In the context of HCI, computers did not exist at the time humans evolved. When interacting with computers, present-day humans apply cognitive strategies they developed many thousands of years ago. Therefore, there is a challenge about how technologies can be designed to fit with the human mind and cognition that may have been outdated today.

2.2 Cognitive Biases in Human-Computer Interaction

The field of Human-Computer Interaction (HCI) was established to investigate the design and use of computer technology. As pioneers of HCI research, Card, Moran, and Newell [25] outlined how the hardware and software of information-processing machines can be designed to optimise the interaction between men and machines by applying psychological sciences. They proposed the human processor model, which drew an analogy of humans with computational processors: when interacting with machines, humans employ perceptual, cognitive, and motor systems to perceive external stimuli, process information, and react accordingly. While the human processor model was initially used to estimate a user's reaction time, HCI researchers employ the model to explain that users employ reasoning when interacting with machines. Importantly, they emphasised the role of **mental models** that shape users' expectations of systems and the user experience.

2.2.1 Mental Models

Pioneer of User Experience Design Donald Norman wrote in his book "The Design of Everyday Things" that humans form mental models when interacting with computers [125]. Rooted in psychology, mental models are our internal representations and simple explanations of the real world that individuals use to understand, interpret, and navigate the world [83]. These models are shaped by an individual's prior experiences and existing beliefs. For example, users have built a strong mental model for a *back* button on a browser - they would understand that clicking on this button will bring them back to a previous webpage. Mental models guide how users predict how a system will work and, therefore, the way users interact with systems. Computing systems have traditionally been designed to adapt to users' mental models for a better user experience. For example, online shopping platforms, computer desktops, and the recycle bin are made to resemble physical shopping experiences, traditional desk workspaces, and recycle bins in the real world. On the contrary, system designs that do not conform to the user's mental model imply that the user's expectation does not align with the way the system actually works. Therefore, the mismatch between the system design and the user's mental model may lead to inaccurate predictions, errors, and confusion, hindering the user experience.

When navigating the real world, cognitive biases arise and can influence how we form mental models. Humans employ heuristics to process information under their cognitive constraints efficiently. These heuristics systematically distort our perception of the external reality and lead to the formation of what Tversky and Kahneman [178] called *subjective reality*. Therefore, heuristics shape how we construct mental models. When interacting with computing systems, users not only form their mental models but also employ heuristics to help fit the system perception into their subjective worldview. As a result, cognitive biases surface in HCI when users form mental models and rely on heuristics to make sense of the systems they interact with [127]. Especially when the user's mental model does not align with the system design, users may calibrate their mental models to align with the systems they interact with. This introduces cognitive challenges to users and gives rise to the application of heuristics and cognitive biases to help them quickly process, filter, and make sense of new knowledge. In the view of evolutionary psychologists, cognitive biases emerge due to the mismatch between the design of the human mind and the present environment [131]. Such mismatches are also present in the design of computing systems that may not align with the user's mental model and, therefore, impose conditions that *trigger* cognitive biases.

2.2.2 Impact of Cognitive Biases in HCI

Cognitive biases, by nature, influence the user behaviour when they interact with computing systems. The HCI literature shows that these biases introduce unintended effects in the user-system interaction. Anchoring bias prompts users to stick to the first information presented to them, for example, top results in search

engines [8], prior information about the AI performance [127], or a default evaluation of a system [150]. Confirmation bias causes users to seek predominantly information that confirms their beliefs. For instance, search engine users mostly clicked on items that support their hypotheses in hand [137]; users tended to like system recommendations that suit their preferences [152]; or people recalled better title headlines that align with their beliefs [102].

HCI research aims to design, build, and evaluate user-facing interventions. One of them is to induce user behaviour change. Hekler et al. [72] suggest that the understanding of cognitive biases can inform the design of behaviour change technologies. Cognitive biases specifically can be used to steer the user behaviour. A well-known concept of *nudging* shows that we can subtly alter the environment in which users make decisions to influence their behaviour [175]. Behavioural scientists Richard Thaler and Cass Sunstein, who introduced the nudge theory, believe that nudging taps into cognitive biases to steer the user behaviour without their awareness. Caraban et al. [23] document different ways to design nudging technologies and chart cognitive biases they tap into. These nudges can be used to promote positive, meaningful change of behaviours, like encouraging healthy food diet, offering alternative choices, or improve password security. However, the same technique can be exploited to push people's behaviour in a manipulative way - or what Thaler called *sludges* [174]. For instance, adding a countdown timer on an online shopping platform prompts users to expedite their decisions. The practice of using technologies to manipulate and deceive users is well-known in HCI as *dark patterns* or deceptive design patterns. Mathur et al. [116] outline different cognitive biases that dark patterns leverage to trick users towards a questionable goal, e.g., subscribing to a service, sneaking an item into a shopping basket, or paying a hidden fee. In sum, the literature supports the idea that computing systems can be deliberately designed to trigger the user's cognitive biases and influence their subconscious, real-world behaviour.

Elements of computing systems, algorithms, and user interfaces can spark cognitive biases, either in a deliberative or unintended manner. These biases can cause positive effects as well as destructive consequences. Therefore, it is important to understand how, what, and when cognitive biases occur in the interaction between users and systems. However, we do not have sufficient understanding of cognitive biases in HCI. While cognitive biases have increasingly been discussed in the HCI community [16, 35, 36, 37], limited research has systematically addressed this research issue. Therefore, this thesis sets out to chart HCI research around the notion of cognitive biases, identify research gaps, and enrich the understanding of the effects of cognitive biases in HCI. Chapter 3 extends beyond the reviewed theoretical concepts of cognitive biases and contextualises it into the HCI literature: we survey the HCI literature and scopes the research around the issue of cognitive biases, and identify the user-system interaction narratives where cognitive biases are present and points out different angles HCI researchers, practitioners, and designers can engage with these biases.

2.3 Summary

Humans have evolved to economise their cognitive processing when navigating in the real world. The same analogy applies when humans interact with computers. Users apply mental shortcuts and heuristics to effectively process information presented by the user interface. Cognitive biases emerge in the interaction as systematic side effects that can influence the way users form perceptions, decisions, and behaviour. In this chapter, we provide an overview of the literature in psychology and behavioural science that reflects how the notion of cognitive bias is defined. Specifically, we highlight the research gap in HCI with respect to the issue of cognitive bias, guiding the grand research question of this thesis: **How do cognitive biases manifest in HCI**? This thesis breaks it down into four research questions presented in Table 1.1. In Chapters 3, 4, 5, and 6, we present the original research contributions of this thesis – which address these questions – and highlight how each chapter builds up on prior work, and discuss the methodology.

Chapter 3

Mapping Cognitive Bias Research in Human-Computer Interaction

3.1 Introduction

People form mental models when interacting with the physical world. Therefore, computing systems and user interfaces are designed to adapt to the users' mental models. Cognitive biases, however, influence how users form such models and interact with computers. Systematic biases and effects have increasingly been studied in HCI, mainly how they help designers account for the design of computing systems. In other words, HCI researchers employ the notion of cognitive bias as a human factor in human-computer interaction. Due to the multidisciplinary nature of HCI, however, researchers engage with cognitive biases from different focuses, contexts, and methodologies. Current research lacks a unified, holistic understanding of how cognitive biases manifest and contextualise in HCI.

In this chapter, we conduct a scoping review to chart how HCI researchers engage with cognitive biases. We search and screen published articles from the HCI literature that study cognitive biases. Based on our corpus with 127 articles, we perform open coding and derive study focuses (Investigating Effects, Mitigating, Observing, Utilising, and Quantifying) and application contexts (Information Interaction and Recommender Systems, Human-AI Interaction, Visualisation, Usability, Behaviour Change, Computer Supported Cooperative Work and Social Computing, Human-Robot Interaction and Autonomous Systems, and Games) of cognitive bias research in HCI, and document the appearances, terminologies, and definitions of cognitive biases in the literature.

Our analysis reveals (1) the narratives of how HCI researchers engage with cognitive biases, (2) the research conduct, and (3) the gap in the literature. Based on the study focuses, we discovered three narratives (A, B, and C) that reflect the dynamics of cognitive biases in users, systems, and designers:

A. Computing systems can trigger as well as mitigate cognitive biases in people;

B. Designers of computing systems capitalise on users' cognitive biases to steer behaviours;

C. HCI researchers develop tools and methods to closer observe cognitive biases.

We further construct a high-level narrative of how HCI researchers engage with cognitive biases, as shown in Figure 1.1. Because computing systems can trigger cognitive biases in users, the interaction between humans and computers can cause (or exacerbate) real-world concerns (e.g., misinformation, trust in AI, or dark

patterns). Motivated by such concerns, HCI researchers build tools and methods to quantify cognitive biases in the interaction to closer investigate and understand the effects of biases. With a better understanding of how cognitive biases manifest in the interaction, HCI researchers derive design considerations which mitigate the undesired effects, leverage biases for good, and address real-world concerns.

We describe the methodology of the scoping review and analysis, as well as their implications in the attached publication, Article I.

3.2 Article I

Nattapat Boonprakong, Benjamin Tag, Jorge Goncalves, and Tilman Dingler. 2025. How Do HCI Researchers Study Cognitive Biases? A Scoping Review. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3706598.3713450

How Do HCI Researchers Study Cognitive Biases? A Scoping Review

Nattapat Boonprakong
School of Computing and Information Systems
University of Melbourne
Parkville, Victoria, Australia
nboonprakong@student.unimelb.edu.au

Benjamin Tag
School of Computer Science and Engineering
University of New South Wales
Sydney, New South Wales, Australia
benjamin.tag@unsw.edu.au

Jorge Goncalves
School of Computing and Information Systems
University of Melbourne
Melbourne, Australia
jorge.goncalves@unimelb.edu.au

Tilman Dingler
Industrial Design Engineering
Delft University of Technology
Delft, Netherlands
t.dingler@tudelft.nl

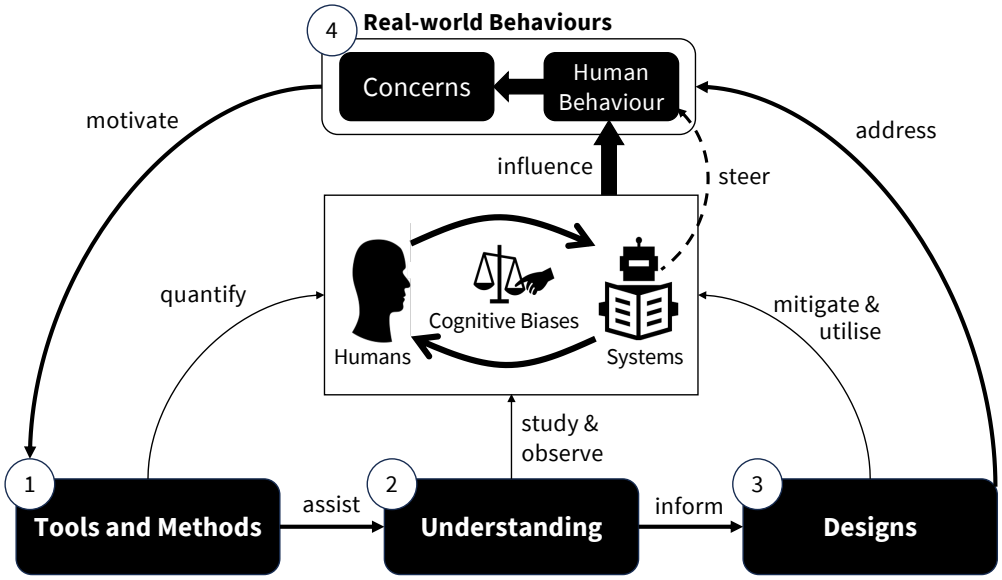


Figure 1: A summary of how HCI researchers study cognitive biases. Computing systems can trigger cognitive biases in humans and influence (or steer) their behaviours and decision-making. Cognitive biases affect the real-world behaviours of humans, which motivates HCI researchers to develop (1) tools and methods measuring the occurrences of cognitive biases to study their effects on the interaction between humans and computers. Consequently, (2) the understanding of cognitive biases informs (3) the design of computing systems, which mitigates or utilises cognitive biases and helps address (4) the real-world behaviour of humans.

Abstract

Computing systems are increasingly designed to adapt to users' cognitive states and mental models. Yet, cognitive biases affect how humans form such models and, therefore, they can impact their interactions with computers. To better understand this interplay, we conducted a scoping review to chart how Human-Computer Interaction (HCI) researchers study cognitive biases. Our findings show that computing systems not only have the potential to induce and amplify cognitive biases but also can be designed to steer users' behaviour and decision-making by capitalising on biases. We describe how HCI researchers develop algorithms and sensing methods to detect and quantify the effects of cognitive biases and discuss how we can use their understanding to inform system design. In this paper, we outline a research agenda for more theory-grounded research and highlight ethical issues when researching and designing computing systems with cognitive biases in mind as they affect real-world behaviour.

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713450

CCS Concepts

• **Human-centered computing → HCI theory, concepts and models**.

Keywords

cognitive bias; decision-making; bias-aware systems

ACM Reference Format:

Nattapat Boonprakong, Benjamin Tag, Jorge Goncalves, and Tilman Dingler. 2025. How Do HCI Researchers Study Cognitive Biases? A Scoping Review. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3706598.3713450

1 Introduction

When interacting with computers, humans form mental models – internal representations of the external reality – based on what they believe, prefer, and are familiar with [30]. The design of everyday user interfaces, such as desktops, digital games, or online websites, is predominantly based on the mental representation of humans from real-world physical objects. However, human mental models are subject to bounded rationality [166]. Humans use simple rules of thumb, developed through their beliefs and experience of the world, to sift through the complexity of everyday decision-making. The pioneers of behavioural economics, Tversky and Kahneman [97, 187, 188] coined such phenomena as **cognitive bias** and documented different ways such biases systematically skew human behaviours and judgements. For example, the anchoring bias makes us tend to rather stick with the first piece of information we encounter [188], or the framing effect influences people to make decisions differently based on how the choices are presented [189]. While computing systems are built to adapt to people's cognitive states and mental models [22, 30], cognitive biases affect how they form such models and, therefore, impact the interaction between humans and computers.

More importantly, cognitive biases can cause harm and open the door to manipulation. Misinformation triggers confirmation bias in Internet users (tendency to seek information that only aligns with one's own beliefs), which lets them believe and propagate such information [60, 164]. Dark patterns [132] and social engineering [23] exploit people's cognitive biases and steer their decision-making. The issue of cognitive biases, hence, becomes a crucial research agenda in HCI to design systems that not only take cognitive biases into account but also remediate their adverse effects [5, 18, 46, 47, 125, 200, 213].

Due to HCI's multidisciplinary nature, research about cognitive biases in HCI is scattered and targets the issue from different angles, methodologies, and application areas. However, there exists no comprehensive review of cognitive bias studies in HCI. In this paper, we provide a scoping review of 127 articles that study cognitive biases in HCI. Our goal is to form a systematic understanding of how cognitive biases manifest in the interaction with computers. Therefore, we analyse the literature and chart how the HCI community conducts research around cognitive biases by categorising papers based on their study focus and application context.

Our results show that HCI research considers cognitive biases as a human factor. HCI researchers aim to understand how human-computer interactions reinforce cognitive biases to inform design considerations that address these biases. We found five different ways HCI researchers engage with cognitive biases (Investigating Effects, Mitigating, Observing, Utilising, and Quantifying) and mapped them in an overarching picture of how HCI researchers study cognitive biases (Figure 1). Computing systems can trigger cognitive biases, which influence, and sometimes steer, real-world human behaviours. Motivated by these concerns, HCI researchers develop **(1) tools and methods** that quantify and capture the occurrences of cognitive biases. These tools and methods help researchers to closer investigate **(2) their effects** on the interaction between humans and computers. With a better understanding of cognitive biases, HCI researchers design **(3) systems and interfaces** that mitigate and leverage cognitive biases, ultimately addressing **(4) real-world human behaviours**. Additionally, we found that cognitive biases are a double-edged sword: not only can their effects steer human judgements, but they are also leveraged for the greater good.

In sum, this paper provides the following contributions:

- We provide a *scoping review* of cognitive bias studies based on a corpus of 127 HCI papers published between January 2010 and May 2024.
- Based on open coding, we derive five ways HCI researchers engage with cognitive biases (Investigating Effects, Mitigating, Observing, Utilising, and Quantifying) and eight application areas where cognitive biases are studied in HCI (Information Interaction and Recommender Systems, Human-AI Interaction, Visualisation, Usability, Behaviour Change, Computer Supported Cooperative Work and Social Computing, Human-Robot Interaction and Autonomous Systems, and Games).
- We map out recommendations and future opportunities for the HCI community to research cognitive biases, voicing the need for community standards, methodological frameworks, and theory-oriented research while discussing the ethical considerations regarding cognitive bias research in HCI. We identify gaps in the literature (Table 2), which guide opportunities for future work.

2 Background

Bias refers to a systematic deviation from the *norm*. There can be several kinds of bias depending on how we set the norm, actors, and application contexts. For example, algorithmic bias describes systematic errors in computing systems that create unfair outcomes [91], and gender bias implies a systematic difference of treatment of one gender over another [41]. In this paper, we focus on *cognitive bias*, first coined by Tversky and Kahneman [188], to refer to a systematic deviation in judgement from the norm of rationality. In this section, we incorporate the literature in cognitive and behavioural science to discuss the notion of cognitive biases, the dual-process theory, and techniques to debias people. We wrap up this section with a discussion of how cognitive biases impact HCI, a review of relevant surveys, and a statement of our contribution to the HCI community.

2.1 Cognitive Biases and Their Interpretations

Humans are not always rational because their cognitive capacity is limited. During the decade of 1950s, Herbert Simon coined the concept of **Bounded Rationality** to explain that, given the complexity of the world and constraints on time and cognitive resources, humans apply **Mental Shortcuts** to faster sift through information and make judgments [166]. Such mental shortcuts can lead to flawed and suboptimal decision-making. Two decades later, Tversky and Kahneman extended this concept and proposed the notion of **Cognitive Bias**: humans employ **Heuristics** as mental shortcuts which systemically deviate their behaviour and the decision-making outcome from the *norm of rationality* [188]. Guaranteeing a fast but suboptimal outcome, heuristics are rules of thumb that humans have adopted through basic instinct, preexisting beliefs, and prior experiences [87]. Through a series of empirical experiments [95, 98, 187–189], Tversky and Kahneman showed that humans employ heuristics and, thus, exhibit several kinds of cognitive biases which systematically skew their decision-making. For example, the anchoring bias makes people rely heavily on the first information presented to them [188]; and the framing effect causes individuals to react to a piece of information differently depending on how it is presented [189]. Subsequent works in cognitive psychology and behavioural science discovered more variations of cognitive biases, such as confirmation bias [142], the halo effect [12] (tendency to rate attractive individuals more favourably for their characteristics), the fundamental attribution error [160] (tendency to overattribute the behaviour of others based on their characteristics), or the Dunning-Kruger effect [113] (tendency to overestimate one's ability despite lacking competence). Up to date, there have been over 180 documented forms[1] of cognitive bias [58]. Meanwhile, recent research has argued that most cognitive biases can be simplified to a form of confirmation bias [147].

Being the by-product of mental shortcuts in decision-making, Kahneman and Tversky viewed cognitive biases as erroneous responses or mental fallacies resulting from the deviation from the norm of rationality [54, 64, 188]. However, recent discourses in psychology have started to shift away from the original interpretation of cognitive biases. The prominent psychologist Gerd Gigerenzer has been a critic of the idea that humans are *biased*, disputing that the norm of rationality does not always exist [194]. In his later works, he argued that heuristics are *fast and frugal* reasoning that helps people make a fast and rational decision at the same time [62, 63]. Some psychologists define cognitive biases as the behavioural consequence of the unconscious, unintended use of mental shortcuts [149, 206]. Hilbert [86] proposes that cognitive biases can be statistically modelled as a result of humans' noisy memory and information processing. In evolution psychology, Haselton and colleagues [78, 79] discuss cognitive biases as rather *design features* of the human mind, citing that humans develop many cognitive biases and heuristics as part of their survival and the natural selection. In this paper, we refer to the more inclusive definition: we consider cognitive biases an **inherent human factor** that broadly and systematically affects and distorts their behaviour.

[1]Not all forms of cognitive bias have been empirically validated in peer-reviewed studies.

2.2 Dual-Process Theory and Debiasing

Human cognition employs two systems of information processing: a fast, automatic, and error-prone **System 1 Thinking** and a slow, conscious, and deliberative **System 2 Thinking** [172, 201]. The so-called **Dual-Process Theory** explains that most of the time, humans resort to using System 1 thinking to make judgments. Psychologists argue that cognitive biases largely emerge from the activation of System 1 thinking [31, 56, 95, 171]. People, therefore, employ heuristics and cognitive biases when sifting through complex information without explicit awareness.

Research in psychology and behavioural science suggests that cognitive biases could be reduced or avoided if people bypass System 1 and shift to System 2 thinking [124, 168, 172]. This can be done through **Cognitive Aid** to guide people to make alternative, rational decisions. Kozyreva et al. [112] propose three main approaches to intervene users away from cognitive biases: nudging, technocognition, and boosting. Nudging [181] changes the environment (i.e., the user interface) and shifts people's behaviour in a subtle way. Notably, nudges can substitute people's autonomous choices with preselected rational decisions. Technocognition [119] are psychological interventions that safeguard people from their biases. For example, slowing down decision-making invites people to reflect on their judgment [170]. Boosting [85] fosters users' metacognition and critical thinking skills to empower control over their decision-making. This also includes education and psychological innoculation [40] that build people's resilience to fast and error-prone thinking.

Correcting people's cognitive biases is, however, not straightforward. Lilienfeld et al. [124] suggest a number of factors that are potential barriers to debiasing. For example, individual differences in working memory and intelligence influence the motivation to engage in rather System 1 or System 2 thinking and, therefore, an individual's receptibility to debiasing interventions [55, 171]. In addition, interventions may not work in the long term and over different contexts [205]. There is also a possibility that interventions may backfire and rather exacerbate the user's existing cognitive biases [207].

2.3 Impact of Cognitive Biases on HCI

Cognitive biases profoundly affect user behaviours, especially when they come into contact with computing systems. The role of cognitive biases influencing the interaction has been discussed in the HCI community over the recent decade [5, 6, 18, 20, 125, 213]. As a result, we observe a growing number of HCI studies investigating cognitive biases (Figure 3). HCI research generally studies the practical aspects of cognitive biases to optimise the human-computer interaction, e.g., how does anchoring bias affect people when they use AI to make decisions [76, 143] or how could nudging interfaces mitigate confirmation bias when people search information on the web [123, 158, 159]. Additionally, interface designers have employed nudges [181], which harness cognitive biases, to steer the user behaviour [28, 109]. Dark patterns [132, 133] and social engineering [23] are interesting case studies where people's cognitive biases are exploited to manipulate their decision-making. Recent works have discussed the notion of *Bias-Awareness* [17, 126]

as computing systems take users' existing cognitive biases into account, mitigate their drawbacks, and maximise their benefits.

2.4 Related Surveys

Our review builds upon the existing surveys in the HCI community that ground on cognitive biases, particularly in the areas of behaviour change technologies and human-centred AI. Hekler et al. [83] survey studies on behaviour change published at CHI between 2002-2012. The authors recommend that insights into cognitive biases in behavioural science can inform the design and evaluation of technologies that help people change their behaviours (e.g., eating more healthy food). More importantly, they suggest that HCI and behavioural science literature remain largely siloed within the two communities. On the other hand, the domain of HCI is in a unique position to contribute to the field of behavioural science. Not only can HCI leverage insights in behavioural science, such as cognitive biases, but it can also complement them through ubiquitous sensing, fast prototyping, and data-driven practice. In the same vain, Pinder et al. [152] provide a critical review of digital health behaviour change interventions and suggest that cognitive biases can be leveraged to induce change in health behaviours (e.g., to quit smoking or reduce anxiety) through the use of Cognitive Bias Modification (CBM), which modifies people's subconscious mental shortcuts by gradually modifying attentional bias (tendency to prioritise attention on a certain type of stimuli), approach bias (tendency to approach rather than avoid repetitive cues), and the priming effect (an individual's exposure to one stimulus influences how they respond to a subsequent stimulus). Caraban et al. [28] review HCI studies that proposed and employed technology-mediated nudges to induce behaviour change. They cover 23 mechanisms of nudging by tapping into people's cognitive biases. While nudges are often criticised as manipulating people's autonomy [156, 176], they find that the majority of the nudges rather transparently promote reflective thinking and influence the user choice rather than implicitly manipulating user behaviour.

Researchers have also explored cognitive biases in AI-assisted decision-making [14, 69, 104, 200]. Wang et al. [200] conduct a literature review of concepts in explainable AI (XAI) and synthesise a conceptual framework of how human cognitive patterns drive the need for building XAI and how XAI can alleviate common cognitive biases. Kliegr et al. [104] review and analyse the effects of cognitive biases on the interpretation of machine learning models. Their work suggests that individual differences (e.g., personality traits and numerical literacy) could influence the effectiveness of cognitive bias mitigation. They also highlight that a study of the effects of cognitive biases should precede bias mitigation to ensure that the bias occurs and the bias mitigation strategy does not backfire. Bertrand et al. [14] survey studies that involve cognitive biases in AI-assisted decision-making. The authors provide an overview of the context in which different cognitive biases affect how XAI systems are designed and evaluated in user studies. The work also outlines that XAI systems can mitigate, as well as cause, trigger, or amplify, the users' existing cognitive biases. The authors argue that not all cognitive biases are harmful. Some of them are inherent to the interaction with the AI explanation.

While the prevailing surveys have explored how cognitive biases manifest in human-computer interaction from different angles of HCI, there is a lack of a comprehensive review of cognitive biases throughout the field. Therefore, we propose a scoping review that charts how the broader HCI community studies cognitive biases. In this regard, we review research papers that investigate cognitive biases in different application domains of HCI and draw a framework (Figure 1) of where cognitive biases are situated in the dynamics of human-computer interaction. To the best of our knowledge, we are the first to comprehensively review the research on cognitive biases across different HCI contexts.

2.5 Our Contributions

Limited work has reviewed the issue of cognitive biases in HCI. While this issue is clearly emerging, research has scattered around different angles and application domains due to the multidisciplinary nature of HCI. Therefore, it is essential to summarise these research spans and assemble the big picture of cognitive biases' prevalence in human-computer interaction. In this paper, we present a *systematic* scoping review that highlights the issue of cognitive biases in HCI. Importantly, **our work provides a holistic overview of cognitive biases in the interaction with computer systems.** We differentiate our work from existing surveys in HCI, which address the question in a specific domain and application scenario (e.g., behaviour change or human-AI interaction). To this end, we augment our discussion with insights from the existing surveys and discussions around cognitive biases in HCI.

Our scoping review seeks to understand the question: **how do HCI researchers study cognitive biases?** Through analysing cognitive bias studies throughout different spaces in HCI, we (1) chart the landscape of cognitive biases in HCI, i.e., what aspects of them are studied and leveraged. This helps us to (2) form guidelines on what HCI researchers should consider when researching cognitive biases. Specifically, our work reflects the practice of how HCI researchers have studied cognitive biases. Furthermore, our work identifies (3) challenges and opportunities for HCI research to develop tools, understanding, and designs that take cognitive biases into account to address concerns about human real-world behaviours. We publish our study corpus and the coding manual as supplementary materials for future work to expand upon.

3 Methodology

Our work qualifies as a scoping review [138]. To address our research question, we systematically examine how cognitive bias research is conducted, identify areas or gaps of research, and clarify the notion of cognitive bias in the HCI literature. To conduct this scoping review, one researcher performed (1) database searches, (2) article screening, and (3) data extraction and coding. Three other researchers iteratively cross-checked the process.

3.1 Database Searches

We followed the PRISMA 2020 guidelines [137] to select relevant publications for this scoping review. First of all, we identified HCI research articles published in leading venues in HCI that are likely to publish work on cognitive biases. This includes venues sponsored by ACM SIGCHI (*e.g.,* CHI, CSCW, TOCHI, CHIIR, or IUI), TVCG,

IJHCS, and IJHCI. We first started with the search query "*cognitive bias*" in publication titles and abstracts to identify research articles relevant to cognitive biases. We then iteratively derived synonyms based on the initial search. Therefore, we identified new terms: "*human bias*", "*confirmation bias*", and "*decision bias*". We note that the terms "human bias" and "decision bias" are relevant to biases related to humans. To maximise coverage of cognitive bias studies, we included "confirmation bias" as a keyword because of its prevalence as the most common form of cognitive bias. Research in psychology suggests that most cognitive biases can be simplified to confirmation bias [147]. Due to the large number of cognitive biases identified in the literature, we found that the terminologies may differ between different articles. More importantly, some papers do not explicitly use the abovementioned terms in the abstract. By title or abstract, we found that searching these terms returned limited results (only 31 results are returned in the ACM Digital Library, as of 31 May 2024). Therefore, we performed **full-text searches** to get better coverage of articles. However, we did not include the generic term "*bias*" as it returned a large number of irrelevant records from the online databases. After fine-tuning, we performed library searches using the following research query:

```
[[Full Text: "cognitive bias" OR "human bias" OR "decision bias" OR "confirmation bias"] AND [E-Publication Date: (01/01/2010 TO 31/05/2024)]]
```

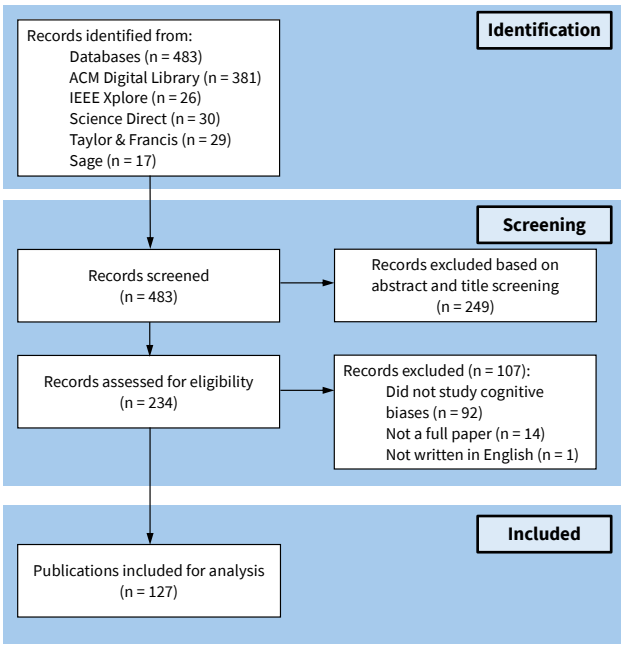


Figure 2: PRISMA 2020 [137] flow diagram for the article screening and selection process.

3.2 Article Screening

We obtained a total of 483 unique records through the database keyword search up to 31 May 2024. We then performed title and abstract screening. At this stage, we included articles that mention keywords that relate to the themes of cognitive biases (including bounded rationality, heuristics, dual-system thinking, decision-making, and mental models). We excluded articles that were clearly outside our scope (e.g., those examining algorithmic bias or media bias). This process left us with 234 papers. Subsequently, we assessed the articles at the full-text level to filter out irrelevant articles that did not investigate cognitive biases. To do so, we excluded papers that did not have cognitive biases as their main focus or study variables (92 papers). In addition, we excluded short papers, posters, late-breaking works, and extended abstracts (14 papers) because they have a different level of maturity than full papers. One paper was excluded because it was not written in English. We derived the final corpus of 127 papers for data extraction. Figure 2 shows a flow diagram for our article selection process.

3.3 Data Extraction and Coding

With our corpus, we created a data extraction sheet to systematically explain our papers from different angles. For each study, we extracted (1) what cognitive biases are studied with their definitions, (2) how cognitive biases are studied, and (3) what application context it covers. We describe the data extraction and coding methodology in the following.

(1) **Cognitive Biases Studied and Their Definitions**: In each paper, we extracted all cognitive biases mentioned with the terms and definitions mentioned by the author(s). We performed a keyword search in the full-text paper to identify possible mentions of cognitive biases. By doing so, we included keywords "bias" and "effect" and cross-checked with Cognitive Bias Foundation's taxonomy of cognitive biases [58], which provides a community-sourced extensive document of more than 180 cognitive biases.

(2) **Study Focus**: For each paper, we extracted how cognitive biases are studied. We then performed open coding and derived five different study focuses in the following bullet points. We note that 14 papers (11.02%) have two study focuses as they consider cognitive biases in multiple angles. For complete information, we publish the data extraction sheet in the supplementary materials.
- **Quantification**: tools, methods, metrics, or mathematical/statistical models to detect, measure, or quantify cognitive biases;
- **Mitigation**: mitigation or prevention of cognitive biases and their adverse effects;
- **Utilisation**: application and utilisation of the effects of cognitive biases in the interaction with computers;
- **Effect Study**: investigation or demonstration of the empirical effects of cognitive biases on the interaction with computers;
- **Observation**: observations or case studies of cognitive biases in people, systems, and their interactions.

(3) **Application Contexts** We extracted the primary application context each paper worked on. Subsequently, we applied open coding to group each paper into eight broad, distinct themes, which represent different areas of HCI research:

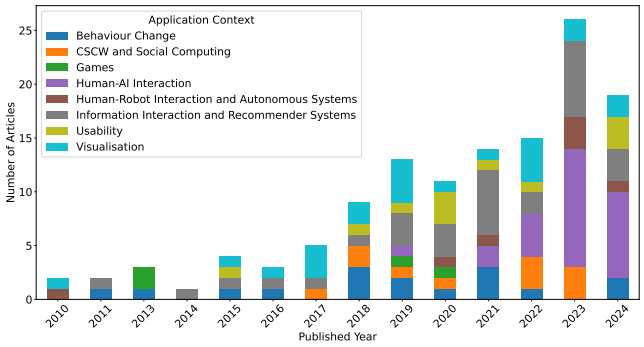


Figure 3: The number of cognitive bias papers by application context and published year, dated from 2010 to May 2024.

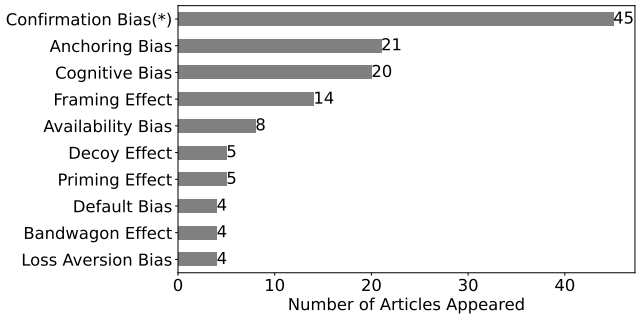


Figure 4: This chart visualises the 10 most frequently mentioned forms of cognitive bias with the number of articles in each of them. Note that some papers may investigate more than one cognitive bias. *The inclusion of "confirmation bias" in the search query may make its occurrences in our corpus more frequent than usual.

(1) Information Interaction and Recommender Systems, (2) Human-AI Interaction, (3) Visualisation, (4) Usability, (5) Behaviour Change, (6) Computer Support Cooperative Work (CSCW) and Social Computing, (7) Human-Robot Interaction and Autonomous Systems[2], and (8) Games.

4 Results

In the following subsections, we provide the analysis of 127 articles in our corpus based on each of our data extraction criteria: publication venue, term and definition of cognitive bias, study focus, and application context.

4.1 Publication Venue and Year

In our corpus, papers published at CHI form the majority of works (40 articles, 31.49%). Other than that, there are papers published at CSCW (12 articles, 9.44%), IJHCI (8 articles, 6.30%), IUI (8 articles, 6.30%), TOCHI (7 articles, 5.51%), CHIIR (6 articles, 4.72%), TVCG (6 articles, 4.72%), and others (40 articles, 31.49%). We observe an upward trend of articles published by year (Figure 3). This reflects that the issue of cognitive biases has increasingly gained attention in HCI research. Notably, roughly half of our corpus (60 articles, 47.24%) was published between 2022 and 2024, with CHI papers making up the majority (27 articles).

4.2 Terms and Definitions of Cognitive Biases

We identified 99 different terms referring to any form of cognitive bias. After merging synonyms (for example, "anchoring bias" and "anchoring effect" are considered the same cognitive bias), we arrived at 92 unique cognitive biases. We found **confirmation bias** to be the most frequently studied (45 articles), which is partially due to its inclusion in the search query. We identified several other cognitive biases, such as anchoring bias (21 articles), the framing effect (14 articles), and availability bias (8 articles). Several papers, however, do not mention any specific form of cognitive biases; for example, 20 articles mention "cognitive bias" or its equivalent terms. Figure 4 charts 10 of the most frequently mentioned cognitive biases in our corpus. We provide the full list of unique cognitive biases in our bias codebook as part of the supplementary materials.

We found discrepancies in the usage of terms of cognitive biases. First of all, *bias*, *effect*, and *heuristics* are used interchangeably. Prominent examples include "anchoring bias" and "anchoring effect" or "availability bias" and "availability heuristics". Goffart et al. [65] cross-termed between "default option effect", "default option bias", and "default bias". We identified many semantically-equivalent terms for "cognitive bias"; for example, "decision bias", "decision heuristics", and "bias in decision making". We also found some cognitive biases to be closely related. There are cross-usages of terms in confirmation bias studies. For example, Liao et al. [123] study *selective exposure* (tendency to focus and seek out information that confirms one's beliefs), which overlaps with confirmation bias. Similarly, Aicher et al. [2] touched on *self-imposed filter bubbles*, a related phenomenon with confirmation bias. Some works also coin context-specific cognitive biases based on existing cognitive biases; for instance, Pafla et al. [151] propose *explanation confirmation bias* based on confirmation bias in explanations provided by AI. We present common cognitive bias synonyms in Table 1.

We discovered some discrepancies in definitions of cognitive bias. Two prominent examples are confirmation bias and anchoring bias. While the majority of papers reference confirmation bias by the definition given in the original work of Nickerson [142], some papers refer to later works in psychology [93] or domain-specific definitions, such as Information Retrieval [5] or Communication Science [106]. Similarly, anchoring bias is mostly defined using the definition in the seminal work of Tversky and Kahneman [188]; however, some papers refer to the definitions in previous HCI research (e.g., [141, 200]). Among articles that refer to the seminal works' definitions, we found variability in the wording. Specifically, many papers frame their cognitive bias definitions as context- or domain-specific. For example, Rieger et al. [158] refer to confirmation bias as "*Users tend to select search results that confirm pre-existing beliefs or values and ignore competing possibilities.*". Naiseh et al. [139] define it as "*Humans favour an XAI classification that is consistent in its output with their beliefs and initial hypothesis.*"

[2]Despite Human-Robot Interaction overlaps significantly with Human-AI Interaction, it deserves a separate category because the physical embodiment of autonomous agents [117].

Table 1: Common cognitive bias synonyms

Cognitive Bias	Semantically Equivalent Cognitive Biases
Cognitive Bias	Decision Bias, Decision Heuristics, Bias in Decision Making, Human Bias
Confirmation Bias	Selective Exposure, Self-Imposed Filter Bubbles, Explanation Confirmation Bias
Anchoring Bias	Anchoring Effect
Availability Bias	Availability Heuristics
Decoy Effect	Asymmetric Dominated Choice
Default Bias	Default Option Bias, Default Option Effect,
Forer-Barnum Effect	Forer Effect, Barnum Effect
Positioning Bias	Positional Bias, Positioning Heuristics
Ambiguity Aversion	Ambiguity Effect
Fundamental Attribution Error	Attraction Effect
Bandwagon Effect	Herd Instinct Bias
Automation Bias	Automation Complacency

Furthermore, we identified 22 instances of cognitive biases without a clear definition stated in the respective papers.

4.3 Study Focuses

We categorised papers in our corpus into five study focuses. Subsequently, these focuses reveal three key narratives of cognitive biases in HCI: (A) computing systems can trigger and mitigate cognitive biases; (B) designers capitalise on cognitive biases in users to steer their behaviours; and (C) HCI researchers develop tools and methods to closer observe cognitive biases. In this section, we review the literature in each narrative based on their study focuses.

A. Computing Systems Can Trigger as well as Mitigate Cognitive Biases in People

4.3.1 Investigating the Effects of Biases. We found 38 papers (29.92%) unveiling cognitive biases that people follow when interacting with computing systems. These studies often set up experiments to demonstrate the effects of cognitive biases, with the goal of understanding cognitive biases as a human factor and deriving design recommendations. Most of the works (30 articles) employed quantitative methods and experimental designs [36, 38, 39, 48, 59, 67, 70, 71, 73, 80–82, 103, 107, 108, 110, 114, 129, 143, 154, 159, 162, 163, 169, 177, 178, 183, 184, 186, 208]. For example, Tomé et al. [186] study loss aversion bias (tendency to avoid losses over achieving equivalent gains) in gameplay and derive design considerations for game designers; Nourani et al. [143] explore anchoring bias in explainable AI and find that users tend to make more errors if they are exposed to system strengths (i.e., they are told that an AI system makes accurate predictions) as an anchor; He et al. [82] investigate the Dunning-Kruger effect on human reliance on an AI system and suggest that people who overestimate their ability tend to rely less on AI advice.

Meanwhile, a limited number of papers study the effects of cognitive biases by using qualitative (2 articles) [61, 173] or mixed methods (6 articles) [37, 43, 50, 68, 134, 144]. For instance, Mendez et al. [134] conduct a qualitative study and a follow-up quantitative experiment to investigate the framing effect in student course selection. Chromik et al. [37] carry out a mixed-method study to examine the illusion of explanatory depth (people's tendency to believe they understand a topic better than they actually do) in explainable AI.

4.3.2 Mitigating Biases. We identified 39 papers (30.71%) that seek to mitigate the effects of cognitive biases. Researchers in HCI employ cognitive aid as a strategy to help users reflect and make rational decisions. More specifically, research proposes tools and user interface designs that serve as cognitive aid [42, 51, 67, 100, 105, 111, 135, 174, 193, 199, 200, 213–216]; for example, Zheng et al. [216] propose an intelligent agent to make the discussion among human teachers more objective and reduce errors in decision-making; and Wang et al. [200] present guidelines for designing explainable AI that encourages its users to avoid amplifying their cognitive biases. Studies also explore using pedagogical tools to teach users critical thinking skills, avoiding their cognitive biases [116, 191, 204]. For instance, Whitaker et al. [204] propose Heuristica, a video game that teaches students to recognise and mitigate cognitive biases using a set of immersive scenarios.

Some research suggests that cognitive biases can be mitigated through system feedback that helps users to reflect on their existing biases [52, 126, 140, 145, 165, 198]. For example, Echterhoff et al. [52] propose a machine learning algorithm that identifies anchored decisions made by users and modifies the presentation order of stimuli to minimise anchoring bias. Narechania et al. [140] develop a visual data analytics tool that shows users their interaction history and encourages them to reflect on their unconscious biases.

We also identified several papers proposing *nudging* [181] to shift users away from biased behaviours. Nudges to mitigate cognitive biases come in the forms of indicators and interface designs [123, 125, 158, 159, 192, 209]. For example, Liao et al. [123]

introduce aspect indicators to reduce selective exposure in information seekers. Rieger et al. [158] employ obfuscation to minimise users' interaction with attitude-confirming search results to mitigate confirmation bias.

Recent works have raised concerns about bias mitigation. Bach et al. [8] propose a list of recommendations for how to incorporate bias mitigation strategies into practical AI applications. Bias mitigation could trigger AI aversion among users and backfire. Therefore, one should consider subtle design patterns, apply bias mitigation periodically rather than constantly, and increase the awareness of potential cognitive biases through the user interface. Some research points out that user-related factors and interaction context could impact the effectiveness of bias mitigation [27, 67, 125, 158, 200]. Graells-Garrido et al. [67] argue that there exists no one-size-fits-all approach to combat cognitive biases. In other words, one bias mitigation strategy does not always work in every individual, context, and scenario. Subsequent studies provide supporting empirical evidence. Cao et al. [27] find that user demographics, such as age and familiarity with probability and statistics, could influence user interaction and, subsequently, amplify the effects of cognitive biases. Rieger et al. [158] point out that situation- and user-related factors (e.g., attitude strength, topic interest, and personality traits) could impact the effectiveness of confirmation bias mitigation approaches.

B. Designers of Computing Systems Capitalise on Users' Cognitive Biases to Steer Behaviours

4.3.3 Utilising Biases. 21 articles (16.53%) leverage cognitive biases to nudge users toward certain behavioural outcomes. We identified two main applications of cognitive biases in the literature: (1) changing user behaviours and (2) incorporating cognitive biases in design. First of all, a significant portion of studies leverage the effects of cognitive biases to shift user behaviours in a predictable way. Lee et al. [118] is among the first to investigate how HCI research could leverage behavioural science to design persuasive technologies. The authors showcase the application of the default bias (tendency to accept what is presented), present bias (tendency to settle for a smaller present reward over a bigger award in the future), and decoy effect (tendency to swap one's preference between two options when a third option is presented) in promoting healthy eating choices. Subsequent studies (e.g., [28, 158, 159, 211, 212, 218]) take the approach of *nudging* [181] – by altering the environment, i.e., the user interface, one can trigger the user's cognitive biases and steer them towards a particular decision or behaviour. For example, Zhang et al. [212] propose an interface nudge to encourage people to reflect on their views on political issues. Zavolokina et al. [211] propose ClarifAI, a tool to nudge users towards more critical news consumption. Meanwhile, some papers take advantage of the effects of cognitive biases to induce behaviour change. Ma et al. [128] employ anchoring bias to promote people's trust in AI. Yamamoto and Takehiro [209] use the priming effect to enhance engagement in critical thinking in web searches. Some research investigates CBM [99, 152], which has been commonly used in psychology to modify people's mental shortcuts towards long-term, habitual behaviour change [94]. Pinder et al. [152] suggest that CBM presents a use-case where cognitive biases are leveraged to change people's habits. Kakoschke et al. [99] propose a CBM-based intervention to reprogram associative links between unhealthy food and automatic appetitive responses, making people eat healthier food.

A notable span of works incorporate cognitive biases in the design of computing systems. Loerakker et al. [127] leverage the framing effect in the design of personal informatics to support self-compassion and positive experiences. Mathur et al. [132] discuss how dark patterns on shopping websites exploit people's cognitive biases and deceive them. Burda et al. [23] investigate cognitive mechanisms of social engineering applications, which manipulate people by triggering their cognitive biases. Theocharous et al. [182] provide a critique of personalised recommendation systems and propose how cognitive biases could be taken into account in these systems.

C. HCI Researchers Develop Tools and Methods to Closer Observe Cognitive Biases

4.3.4 Observing Biases. We found 25 articles (19.68%) that consider cognitive biases as a human factor in the interaction with computers. Some research observe the manifestation of cognitive biases [3, 4, 7, 60, 76, 115, 139, 157, 164]. For example, Rho et al. [157] and Mantri et al. [131] demonstrate the framing effect in user comments in forums of publishers. Haque et al. [76] show that law enforcement agents tended to exhibit anchoring bias when interacting with crime maps presented by decision support systems. Some papers discuss the unintended consequences of cognitive biases arising during the interaction [72, 151, 179]. For instance, Pafla et al. [151] point out the risk of saliency maps triggering confirmation bias when interacting with AI explanations. Habib et al. [72] also suggest that confirm-shaming in cookie consent interfaces (i.e., highlighting negative outcomes of not accepting optional cookies) could target users' loss aversion bias. Moreover, some papers discuss their results from the perspective of cognitive biases and decision-making [32, 164, 185]. Shi et al. [164] study the effect of news veracity on cognitive load. They employ cognitive load as a surrogate of System 2 thinking activation, which links to the manifestation of cognitive biases when processing information. Additionally, we identified a number of survey papers documenting cognitive biases in human-computer interaction, such as behaviour change technology [83], visualisation [44], interactive information retrieval [5], and dark patterns [132].

4.3.5 Quantifying Biases. 18 papers (14.17%) propose methods to detect and quantify cognitive biases. This span of research predominantly sets up experiments that induce cognitive biases and measure cognitive biases using different metrics. A number of studies utilise machine learning algorithms to infer cognitive biases through user interaction data [52, 126, 145, 196, 197]; for example, Wall et al. [196] train Markov models to recognise biased behaviours through the user interaction with scatter-plot visualisation; and Echterhoff et al. [52] use a combination of Support Vector Machines (SVM) and Long Short-Term Memory (LSTM) Neural Networks to capture anchoring bias from sequential decision data. Some papers derived statistical and mathematical modelling as measures of cognitive biases [2, 49, 114, 125, 155]. For instance, Rastogi et al. [155] employ Bayesian modelling for human decision-making in human-AI interaction. In their paper, anchoring and confirmation biases are modelled as scenarios when certain model weights are

high. We also found some papers proposing metrics that are derived directly from the original definition in behavioural science or prior literature [25, 33, 45, 88, 129, 186]. Ma et al. [129], for example, quantify overconfidence bias (tendency to have more confidence in one's own abilities) as the difference between the user's expected accuracy of the model and their self-reported self-confidence.

Furthermore, we identified studies that used sensor data to detect cognitive biases [17, 77]. Harris [77] evaluate the bandwagon effect in relevance judgment by using eye-tracking data. Boonprakong et al. [17] employ Functional near-infrared spectroscopy (fNIRS) and feature engineering to measure the effects of cognitive biases when comprehending different opinions.

4.4 Application Contexts

Based on open coding, we identified eight application contexts that span the research space of our corpus. We briefly review the literature in each application context in the following.

4.4.1 Information Interaction and Recommender Systems. We found the majority of cognitive bias studies to be concerned with the area of Information Interaction and Recommender Systems (30 articles, 23.62%). Interestingly, most studies are around the phenomenon of biased information seeking in people, which includes selective exposure [15, 123, 163, 174], misinformation [60, 100, 164], echo chambers [49], and filter bubbles [2]. Some research investigates user attitudes and viewpoints, as people employ them as a principal heuristic for processing information [16, 17, 45, 50, 51]. A significant portion of papers also work on the issue of recommendation systems as they could amplify and have the potential to exploit cognitive biases in users [67, 162, 177, 182]. For example, Graells-Garrido et al. [67] suggest that algorithms and user interfaces should be used in a combination that helps users avoid cognitive mechanisms that lead to biased behaviours.

4.4.2 Human-AI Interaction. The second most popular area is Human-AI Interaction (26 articles, 20.47%). It is a recent and rapidly growing area of study, with most articles published in 2023 and 2024, as shown in figure 3. These papers discuss the problems of explainable AI [37, 71, 73, 143, 151, 200], AI-assisted decision making [8, 25, 27, 35, 52, 128, 129, 155, 178], and trust-reliance in AI [81, 82, 103, 185]. These papers unveiled different cognitive biases that shape the user's mental model when interacting with AI, such as confirmation bias [27, 71, 151, 200], anchoring bias [8, 128, 143, 155], framing effect [52, 73, 103], or the Dunning-Kruger effect [81]. The literature mentions that not only AI systems could trigger and amplify existing cognitive biases in people [8], but other factors, such as limited time [155, 167] and technological expertise [185], could as well influence and facilitate cognitive biases.

4.4.3 Visualisation. We found 22 articles (17.32%) discussing different aspects of cognitive biases in information visualisation. First of all, a number of articles suggest that cognitive biases impact how users interact with and make decisions based on the visualised information [36, 43, 107, 108, 134, 154, 192]. For example, Kong et al. [108] investigate the effects of confirmation bias on how people recall the visualisation of titles. Mendez et al. [134] suggest that the framing effect in visualisation can induce students to put more effort into course selection. Notably, some research proposes that cognitive biases can be detected and quantified through the user's behaviours [145, 196, 197]. For instance, a series of works by Wall et al. [196, 197] propose computational methods (e.g., Markov models) to characterise and predict cognitive biases from the systematic deviation in user interaction from a theoretical baseline. Studies also discuss how cognitive biases can inform the design of interactive visualisation systems [9, 197, 199] as well as the presentation of information [43, 108, 192, 213] to avoid unintended effects from visualisation. Moreover, some research discusses the symbiosis of cognitive biases and visualisation systems, as we can use either of them to improve the other [140, 145, 197].

4.4.4 Behaviour Change. We identified 16 articles (12.60%) that focused on behaviour change. These papers highly overlap with papers that utilise cognitive biases (12 articles, see Table 2), as they seek to shift user behaviour by tapping into users' cognitive biases. Notably, works by Lee et al. [118] and Hekler et al. [83] pioneer how cognitive biases can be used to induce behaviour change in HCI, for example, to encourage healthy habits, Lee et al. [118] design a webpage for snack buying that shows two healthy food choices on the first page, requiring users to click next to see other food options. This taps into users' default bias and steers their food selection behaviour. Later research discusses cognitive biases in nudges and persuasive technologies [28, 29, 65, 109, 156, 158, 212, 218]. Some research expands the discussion on CBM to induce long-term behaviour change in health-related behaviours [99, 152]. Moreover, some papers discuss the potential of cognitive assistants in boosting reasoning and critical thinking skills in people [116, 211].

4.4.5 Usability. Eleven articles (8.66%) investigate cognitive biases from the angle of usability. We found a significant portion of research discussing how cognitive biases can influence the usability of interactive systems. For example, Veytizou et al. [193] suggest that the halo effect can influence user opinions on usability. A series of works by Mathur and colleagues [132, 133] discuss several cognitive biases (e.g., anchoring bias, bandwagon effect, or default bias) that could be exploited by dark pattern user interfaces. Alqahtani et al. [4] study uncertainties in the interaction with self-tracking systems. They argue that users may rely on heuristics and cognitive biases (e.g., confirmation and availability bias) as strategies to avoid uncertainties in the interaction. Chen et al. [33] investigate the Weber-Fechner law as a cognitive bias that influences the perceived visual consistency when users view visual icons across different devices caused by adaptive scaling.

4.4.6 Computer-Supported Cooperative Work (CSCW) and Social Computing. Eleven articles (8.66%) focus on the interactions beyond an individual. We found a notable portion of papers discuss the impact of cognitive biases in crowdsourcing and collective ratings [3, 34, 75, 88, 184], suggesting that humans (i.e., annotators) have a potential to introduce biases into data and algorithms. For example, Hube et al. [88] and Thomas et al. [184] suggest that different cognitive biases can impact and introduce errors to data annotations. Haq et al. [75] propose a method to mitigate errors from cognitive biases in data workers. In addition, some papers investigate cognitive biases in the context of human collaborative technologies. For instance, Shi et al. [165] suggest that activity

Table 2: Categorisation of cognitive bias papers by study focus and application context with their respective count and references. Note that some papers have more than one study focus, therefore, they can have multiple entries in a row. Meanwhile, papers categorised by application context are mutually exclusive. (n/a means no paper examining in that category)

Study Focus × Application Context	**Mitigation** (N=39)	**Effect Study** (N=38)	**Observation** (N=25)	**Utilisation** (N=21)	**Quantification** (N=18)
Information Interaction and Recommender Systems (N=30)	10 articles [51, 67, 100, 123, 125, 153, 158, 174, 209, 214]	10 articles [50, 61, 67, 70, 144, 162, 163, 169, 177, 208]	4 articles [5, 60, 164, 179]	3 articles [15, 182, 209]	6 articles [2, 17, 45, 49, 77, 125]
Human-AI Interaction (N=26)	7 articles [8, 27, 35, 52, 155, 190, 200]	11 articles [37, 71, 73, 81, 82, 103, 110, 129, 143, 178, 183]	6 articles [26, 76, 139, 151, 167, 185]	1 article [128]	4 articles [25, 52, 129, 155]
Visualisation (N=22)	10 articles [11, 42, 111, 126, 140, 145, 192, 198, 199, 213]	6 articles [36, 43, 107, 108, 134, 154]	3 articles [9, 44, 131]	3 articles [127, 134, 192]	4 articles [126, 145, 196, 197]
Behaviour Change (N=16)	2 articles [116, 159]	2 articles [68, 159]	1 article [83]	12 articles [28, 29, 65, 99, 109, 118, 146, 152, 156, 211, 212, 218]	n/a
Usability (N=11)	3 articles [105, 193, 215]	n/a	5 articles [4, 32, 72, 121, 133]	2 articles [23, 132]	1 article [33]
CSCW and Social Computing (N=11)	5 articles [75, 88, 135, 165, 216]	3 articles [59, 173, 184]	3 articles [3, 115, 157]	n/a	1 article [88]
Human-Robot Interaction (N=7)	n/a	4 articles [38, 48, 80, 114]	3 articles [7, 84, 150]	n/a	1 article [114]
Games (N=4)	2 articles [191, 204]	2 articles [39, 186]	n/a	n/a	1 article [186]

traces can help mitigate cognitive biases in peer evaluations. Zheng et al. [216] point out that incorporating AI in group decision-making can stimulate human members to reflect on their logic and reduce cognitive biases in flawed decision-making.

4.4.7 Human-Robot Interaction and Autonomous Systems. Seven articles (5.51%) discuss cognitive biases in the interaction with robots and/or physical autonomous systems [7, 38, 48, 80, 84, 114, 150]. For example, Paepcke and Takayama [150] find that confirmation bias affects how users set expectations about the robot's ability. Hayashi et al. [80] show that anchoring bias makes people stick with human experts' suggestions for decision-making over those suggested by robots. Some research also investigates cognitive biases when acting with autonomous vehicles [38, 48]. Colley et al. [38] suggest that the presence of autonomous vehicles could trigger the halo effect in pedestrians who signal with the vehicle. Interestingly, we found no work that seeks to mitigate cognitive biases in the domain of human-robot interaction.

4.4.8 Games. Four articles (3.14%) discuss the impact of cognitive biases in gameplay. Constant and Levieux [39] find that dynamic game difficulty adjustment could trigger players' overconfidence bias and illusion of control. Tomé et al. [186] study how lost aversion bias impacts how players make decisions in games. Interestingly, some research suggests games could be incorporated into learning systems to mitigate cognitive biases [191, 204]. Veinott et al. [191] examine how serious video games can improve people's ability to be aware of and, therefore, overcome their own cognitive biases.

5 Discussion

In this section, we reflect on higher-level insights obtained from our analysis of the articles in our corpus. We address the main research question of how HCI researchers study cognitive biases, explain the role of cognitive biases as a double-edged sword in the interaction with computers, and discuss the ethical implications of the exploitation of cognitive biases in users of computing systems.

5.1 How Do HCI Researchers Study Cognitive Biases?

The literature clearly indicates that cognitive biases are pervasive in HCI. Humans are not always rational. They are susceptible to making flawed decisions. When interacting with computing systems, user interfaces have the potential to trigger users' cognitive biases. Subsequently, cognitive biases systematically affect users' mental models and, subsequently, their real-world behaviours when interacting with computing systems. According to our findings in section 4.3, the manifestation of cognitive biases in HCI can be explained in three layers: **(A) systems could trigger or remedy existing cognitive biases in users**; **(B) interface designers capitalise users' cognitive biases to (either intentionally or not) steer and manipulate their behaviours**; and **(C) HCI researchers observe cognitive biases in the interaction with computers and develop tools and methods to closer study them.** Although these scenarios do not always happen simultaneously, they all highlight that cognitive biases are a crucial human factor in designing computing systems and user interfaces.

HCI researchers study cognitive biases in the interaction between humans and computers, not only to understand them as a human factor, but also to inform the design of computing systems to better adapt to the user's mental models. By deriving insights from behavioural economics and psychology, HCI research develops tools and metrics to detect, quantify, and study cognitive biases in human-computer interaction more closely. Moreover, many HCI papers employ cognitive biases as a tool to study human behaviours and decision-making. In line with the recent literature in psychology, HCI researchers treat cognitive biases as features of the human mind [78, 79] and, therefore, incorporate them as a human factor. In sum, insights from cognitive bias studies help the HCI community derive recommendations and practicality for designs that take biases in people into account. Recent research [17, 126, 217] has introduced the notion of *bias-awareness*, which refers to the ability to detect, understand, and take into account cognitive biases in people and computing systems. HCI researchers leverage bias-aware systems to address human behaviour and its associated real-world concerns, such as helping humans make objective decisions [155, 165], calibrating their trust in AI [103, 185], or guiding them how to discern online propaganda [20, 211]. Figure 1 summarises how HCI researchers work with cognitive biases to (1) develop tools and methods, (2) better understand people and their biases, (3) inform the design of computing systems and user interfaces, and (4) address real-world human behaviours.

5.2 Cognitive Biases as a Double-edged Sword in HCI

The prevalence of cognitive biases in HCI is a double-edged sword; there are negative and positive effects arising from cognitive biases. A significant number of articles in our corpus outline how cognitive biases could result in negative consequences, such as undermining the collaboration between human and AI systems [81, 82, 103, 185, 200], facilitating the spread of misinformation [60, 100, 164] and unhealthy information behaviours [15, 123, 163, 174], inducing errors when navigating through information [36, 43, 107, 108, 134, 154, 192], or affecting the quality of crowdsourced data [3, 34, 75, 88, 184]. Subsequently, our review identifies different methods proposed to mitigate the adverse effects of cognitive biases. HCI researchers study and employ *cognitive aid*, as introduced in psychology, to help users reflect on themselves and make informed decisions. Based on the dual-system theory, these systems shift people towards using the slower and more deliberative System 2 thinking rather than the fast and error-prone System 1 thinking to make decisions. We found such cognitive aid comes in the form of either *nudging* or *boosting*. While nudging guides people to shift their behaviour, boosting empowers their cognitive and motivational compentencies [85, 112]. The latter approach, boosting, appears in our review as, for instance, tools to teach users critical thinking skills and how to spot their cognitive biases [116, 191, 204].

On the other hand, cognitive biases can benefit the interaction. Our review identifies different studies and research leveraging cognitive biases for the greater good. One prominent example, which is mentioned above, is nudging. Nudging capitalises on the users' cognitive biases to steer them towards a certain behavioural outcome [28]. At the same time, nudges present a use-case where **different cognitive biases can cancel each other out**. Rieger et al. [158] employ nudges as targeted obfuscation in search results to decrease user interaction with attitude-confirming information. Explained by [28], this nudge triggers status-quo bias, which prevents users from interacting with the obfuscated items, and, therefore, mitigates confirmation bias. Furthermore, cognitive biases can be leveraged to induce long-term behaviour change in the form of CBM [99, 152], which has been largely used in healthcare and intervention-focused approaches (e.g., helping individuals quit smoking, eat healthier, or alleviate anxiety symptoms).

Our findings, therefore, suggest that humans are cognitive misers. Some forms of cognitive bias can present in users and their interaction with computing systems. We recommend that the HCI community designs systems aiming to mitigate the negative effects of biases while considering what benefits we can leverage from the users' existing cognitive biases.

5.3 Ethical Considerations from the Exploitation of Cognitive Biases in People

We must, however, acknowledge the ethical implications arising from exploiting inherent human biases. Humans often exhibit cognitive biases without explicit awareness. Therefore, designs and technologies that harness these cognitive biases could risk manipulating their behaviours. Richard Thaler, who first coined the term *nudges*, discussed that the same techniques used to nudge people could be used for negative intentions – the so-called *sludges* [180]. Nudges, on the other hand, could harm user autonomy by steering their behaviours without their awareness and consent [21, 156, 175, 176]. Daniel Kahneman himself and other psychologists also criticised nudges as potential *benevolent paternalism*: governments or ruling institutions can employ nudges to manipulate individuals' choices by assuming the "best interest" of the people [92, 95, 176]. Dark patterns [132] and social engineering [23, 57] are well-researched practical examples where cognitive biases are misused to influence people's decision-making. Boonprakong et al. [17, 19] argue that the same techniques to detect and mitigate cognitive biases could

be used to reaffirm and steer people's beliefs. The Cambridge Analytica scandal [13] demonstrate that people's attitudes could be derived from social media interaction data and, in turn, used to target and sway their opinion-making.

Designers and HCI practitioners shape user experiences and build systems that steer user behaviour. Therefore, they should be held accountable for the ethical implications arising from their design choices. Designers should be well aware that users are inherently susceptible to cognitive biases and of what harm they potentially cause [19] (e.g., people who can fall victim to misinformation are less likely to be debunked [120]). One solution to address ethical concerns could be promoting transparency, which gives users the awareness of their cognitive biases being used. Bias-exploiting interfaces can practically ask for informed consent from users that their behaviour may be subconsciously steered. Zhu et al. [217] suggest that, by giving the awareness of how systems collect and process data, users can make informed decisions. Moreover, legal restrictions, such as the European Union's General Data Protection Regulation (GDPR)[3], could also limit how much (sensitive) data systems can collect for bias detection and quantification.

6 Recommendations to the HCI Community

HCI is uniquely at the intersection of multiple disciplines. Our review suggests that the HCI community derives definitions and theories from psychological and behavioural sciences. However, there exists a gap between HCI and these fields. In this section, we discuss the need for the community to establish a standard for bias terminologies and to closer engage with behavioural science and psychology.

6.1 The Use of Cognitive Bias Terminologies

Our review maps out a series of discrepancies in the use of cognitive bias terminologies and definitions. Variations from the standard terminology, including the use of self-defined terms, could cause confusion among the readers and those who search literature using certain keywords. We originally derived only a few records (for example, N = 31 on the ACM Digital Library) when performing a keyword search based on title or abstract. For this reason, we extended our search to records from full-text searches, which returned significantly more results. This implies that (terms for) cognitive biases are varied and often only mentioned in the full-text paper. We suggest that authors in the HCI community should (1) clearly mention the cognitive biases they study, (2) give explicit definitions, and (3) provide a connection to the notion of cognitive biases. By providing clarity and connection to psychology and behavioural science, HCI researchers can improve the internal validity of their studies. Nonetheless, some psychologists argue that the field of psychology itself is experiencing a similar problem, as new constructs are redundantly invented for existing psychological constructs [53, 74].

The lack of a clear definition and reference to cognitive biases also poses a problem. While cognitive biases are relatively well-known phenomena, we recommend authors in the HCI community clearly define cognitive biases and provide definitions and references to make sure that their studies are theoretically grounded. We suggest that, if possible, definitions could link to the seminal works in behavioural economics and psychology; for example, linking anchoring bias with the seminal work of Tversky and Kahneman [188] or confirmation bias with the work of Nickerson [142]. Furthermore, with some cognitive biases being made context-specific (e.g., confirmation bias as "Selective Exposure" or "Self-Imposed Filter Bubbles", and the Dunning-Kruger effect as "Illusion of Explanatory Depth"), we suggest that authors should make clear that they study such cognitive biases in a specific context.

We argue that it is necessary to establish a community standard for terminologies and definitions. For example, what is the distinction between *heuristics* and *cognitive biases*? Should we use "Ambiguity Effect" or "Ambiguity Aversion"? More importantly, our findings show that there is a discrepancy in understanding whether cognitive biases, as a human factor, are heuristics people use to make faster decisions or *consequences* from the use of such heuristics. Most papers say "mitigate cognitive biases": does it mean we mitigate the cognitive bias itself or its effects? As the notion of cognitive biases is increasingly discussed in the HCI community, we envision that the community could find a consensus on the best practices to report research regarding cognitive biases.

6.2 Closer Engagement with Behavioural Science and Psychology

Cognitive biases are grounded in the fields of behavioural, psychological, and cognitive sciences. Therefore, our review voices a need for the HCI community to connect with the literature and scholars in these domains. Prior research suggests that insights in behavioural and cognitive sciences can inform and integrate with the HCI field to conduct studies that are grounded in theory rather than relying on intuition [83, 155, 213]. Because our understanding of human decision-making has been limited, we envision that HCI research can complement behavioural science. With the ability to fast prototyping and running user studies (such as A/B testing), HCI researchers can quickly verify behavioural science theories [83]. The field of HCI also offers multidisciplinary perspectives that augment the traditional understanding of cognitive biases. Tomé et al. [186] discuss that, while loss aversion bias has been studied in behavioural science, we have little understanding of how it affects gameplay.

Recent research in psychology has signalled a shift away from Kahneman and Tversky's original interpretation of cognitive biases [63, 79, 203]. Different schools of psychologists (e.g., Kahneman & Tversky [188] vs. Gigerenzer [62]) may view the issue of how humans satisfy their cognitive constraints differently. The notion of cognitive bias, therefore, may not offer the most robust explanation that fits human behavioural effects/phenomena [161, 202]. While the HCI community has widely adopted the traditional notion of cognitive biases as an explainer of HCI-related effects (e.g., selective exposure, information framing, or dark patterns), the field could also consider and keep up with the more recent or inclusive definitions, such as Gigerenzer's *fast and frugal heuristics* [62], the challenge of humans' cognitive limitations [166], or noise in human decision-making [96].

[3]https://gdpr-info.eu/issues/personal-data/

7 Avenues for Future Research

Our review signals multiple paths for future research on cognitive biases in HCI. In this section, we discuss the potential for establishing frameworks to study cognitive biases, considerations of effectively leveraging and mitigating cognitive biases, underexplored application contexts, and improving the external validity of cognitive bias studies in HCI.

7.1 Methodological and Theoretical Framework for Studying Cognitive Biases in HCI

Limited research (18 papers, 14.17% of corpus) explores methods, tools, and frameworks to quantify cognitive biases in the interaction with computers. Our findings show that different papers pursue different approaches to quantifying cognitive biases through metrics, statistical modelling, and physiological sensors. Yet, these methods tend to be catered specifically to particular cognitive biases (e.g., anchoring bias) or application contexts (e.g., interaction with information visualisation or AI-assisted decision-making). We suggest that there could be quantification tools that are agnostic to particular cognitive biases or scenarios. Future research could also explore tools to indicate cognitive biases as simply a *deviation from the norm of rationality*. Similarly, Liu [125] proposes a probabilistic framework for human fairness in decision-making. Boonprakong et al. [17] investigate physiological expressions of *any* cognitive biases in opinion comprehension.

7.2 Considerations of Effectively Leveraging and Mitigating Cognitive Biases in HCI

7.2.1 Leveraging Cognitive Biases. Limited works have explored how cognitive biases can be harnessed for the greater good. Our understanding of cognitive biases in HCI is emerging; therefore, we envision the HCI community could avoid the harm of cognitive biases and leverage their benefits. From the literature, we identify two ways that cognitive biases can be leveraged. Firstly, by understanding how cognitive biases affect human behaviours, HCI researchers can take cognitive biases into account when designing interactive systems. Cognitive biases can be applied to induce systematic behaviour changes for the greater good, such as critical thinking engagements [209, 211, 212], healthy food diet [99], or support self-compassion [127]. Secondly, cognitive biases are used to steer users' behaviour in the form of nudges. Caraban et al. [28] document 23 different mechanisms of nudging and discuss cognitive biases associated with each type of nudge. However, research in behavioural science suggests several shortcomings of nudges. Specifically, the effectiveness of nudges may be limited and not sustained over time [10, 21]. At the same time, a plethora of individual and contextual factors may positively or negatively influence the occurrences of cognitive biases [19, 27, 158] and, subsequently, compromise the success of nudges. We, therefore, argue that future research could (1) conduct longitudinal studies to evaluate the effectiveness of nudges and (2) investigate how we could consider individual and contextual factors in utilising cognitive biases effectively. Ultimately, by leveraging the effects of cognitive biases and avoiding their harm, we could enhance the capability of humans and optimise the symbiosis between humans and machines.

7.2.2 Mitigating Cognitive Biases. Research in HCI has pointed out how cognitive biases can negatively affect the interaction with computers. In response, various studies also propose methods to alleviate these effects. However, our findings suggest many research gaps, which echo the discussions in psychology and behavioural science regarding barriers to effective debiasing [124, 205]. First of all, some bias mitigation techniques, such as providing system feedback or pedagogical tools, have been evaluated in only certain contexts like visualisation and human-AI interaction. Future research could assess the effectiveness of the same set of techniques across different application contexts like information interaction or gameplay. For example, Narechania et al. [140] suggest that pointing users to their interaction history in a visual analytics tool could help users reflect on their existing cognitive biases. However, the question remains whether the same approach works if social media users are presented with browsing history.

Secondly, limited research has investigated whether bias mitigation works in practice. Bach et al. [8] suggest a set of recommendations when incorporating bias mitigation into real-world applications. Recent research in behavioural science has also discussed the danger of the *backfire effect*, which could unexpectedly overturn the effectiveness of an intervention [1, 24]. Future research could investigate when and how the backfire effect occurs when mitigating cognitive biases.

7.3 Understudied Application Contexts

Our findings indicate a number of areas with limited research. Table 2 suggests research fixation and gaps. Based on application context, we found that limited research has investigated the contexts of behaviour change, usability, CSCW, human-robot interaction, and games. Research around behaviour change has predominantly focused on utilising cognitive biases. Meanwhile, limited research seeks to quantify the effects of cognitive biases when steering user behaviour. Most research in our corpus considers one side of the picture – either quantifying either the effects of cognitive biases or the effectiveness of behaviour-change interventions – assuming that cognitive biases take effect regardless of the individual and interaction contexts. With the ability to quantify the effects of cognitive biases, one can empirically measure to what extent people's behaviour has changed and how strong the effect is. Similarly, some papers in our corpus showcase how the angles of bias mitigation and quantification can be harmonised [52, 88, 125, 126, 145, 155]. We envision that future research could consider multiple angles to closer study cognitive biases.

We found limited research in the realm of usability that considers cognitive biases, although this issue is central in HCI research. We suggest more research could explore interface design elements that trigger and reinforce cognitive biases (e.g., [132, 136] discuss how dark patterns are connected with certain cognitive biases). Also, no research in our corpus discusses creativity in conjunction with cognitive biases, such as the issue of design fixation [90] – a cognitive bias that makes people stick to a set of pre-conceived ideas and restrict the choices of design. Some HCI scholars [102, 195] have empirically investigated design fixation, however, they make minimal connection with the discourse around cognitive biases.

Additionally, future research could consider cognitive biases beyond just humans and systems, specifically in the domains of CSCW and human-robot interaction. While some works employ bias mitigation strategies in human collaborative work [165, 216], limited research has explored where these biases come from and the potential to leverage them. Future research may address the question of how do computing systems systematically trigger cognitive biases in a crowd of users (e.g., human teams or social network users), and how cognitive biases can be leveraged in coorperative tasks.

7.4 Expanding External Validity

There are a number of threats to the external validity of existing cognitive bias studies in HCI. First, most studies are conducted in a controlled environment. Outside of the laboratory, a plethora of external factors could affect the way people exhibit cognitive biases. We suggest future research consider running in-the-wild studies to reflect how cognitive biases manifest in real-world interactions. In addition, only a few papers consider multiple forms of cognitive bias in conjunction. Humans could exhibit more than one cognitive bias at the same time (e.g., [213]), while multiple cognitive biases can interact, reinforce, or cancel each other [28, 213]. Future research could conduct studies that consider possible cognitive biases that could occur and confound the study variables. For example, studying confirmation bias in social media browsing might introduce anchoring bias when viewing the contents in a sequence and overconfidence bias when the user has more expertise in the content topic. With the awareness of potential cognitive biases in an experiment, researchers could consider ways to minimise these confounds, such as counterbalancing (anchoring bias) and taking topic expertise as a control variable (overconfidence bias).

8 Limitations

This paper has several limitations. First of all, our corpus does not exclusively cover all cognitive bias studies in HCI. We believe, however, that the inclusion of SIGCHI-sponsored venues (such as CHI, CSCW, IUI, or CHIIR) gives a representative view of the HCI community's discourse around cognitive biases. Moreover, the choice of search keywords may not cover all cognitive biases. The literature refers to cognitive biases in many different ways [14]. It is possible that some papers investigate a relevant issue around cognitive biases but do not explicitly mention the term cognitive biases, for example, bounded rationality [101, 130, 148], decision-making fairness [66, 210], systematic bias [89], design fixation [195], self-selection bias [217], or selective exposure [122]. We share the same sentiment with Kliegr et al. [104], who argue there is an abundance of cognitive phenomena that are not regarded as cognitive biases. We also acknowledge that, since the definition of cognitive biases and heuristics (according to Tversky and Kahneman's original school of thought [188]) has been challenged by many psychologists (e.g., [62, 79, 96]), HCI studies may move away from such concepts and use other terms. While our scoping review mainly investigates the use of cognitive biases in HCI research, we reflect that the actual literature around cognitive biases utilises a diverse range of terms that may not be included in our review.

The list of cognitive biases studied and their figures (Figure 4) may be subject to discussion and change as the research landscape is evolving and some cognitive biases could be considered as a specialised form of another cognitive bias. For example, recency bias (the tendency to more easily remember what happened recently) is considered a form of the peak-end rule. To the best of our knowledge, there has been no commonly agreed-upon taxonomy for cognitive biases, with the taxonomy of the Cognitive Bias Foundation [58] providing extensive coverage of more than 180 cognitive biases. Additionally, we acknowledge that the results could be subject to screening biases and personal views as only one researcher performed article screening and coding.

9 Conclusion

Humans employ heuristics and mental shortcuts to effectively make decisions under their inherent limited cognitive capacity. These shortcuts result in cognitive biases, which systematically influence how humans interact with computers. The HCI community has increasingly discussed this issue in the recent decade. This scoping review charts how HCI researchers study cognitive biases. From 127 articles identified, we found that the prevalence of cognitive biases in HCI gives opportunities for researchers to study, mitigate, and leverage their effects to inform designs, optimise the interaction, and address real-world human behaviour. The literature suggests that cognitive biases are a two-edged sword. While we can leverage their effects to induce behaviour change, the same mechanism can be used to manipulate people's decision-making and harm their autonomy. Our results reveal various terminologies and definitions for cognitive biases, suggesting a lack of standards for terming and defining cognitive biases in HCI. To this end, our findings promise several avenues for future research to better understand cognitive biases in the interaction between humans and computers and the need to connect with the literature in behavioural science and psychology.

Acknowledgments

We thank the anonymous reviewers and members of the University of Melbourne's HCI group for their suggestions, which positively helped shape this scoping review.

References

[1] Zhila Aghajari, Eric P. S. Baumer, and Dominic DiFranzo. 2023. Reviewing Interventions to Address Misinformation: The Need to Expand Our Vision Beyond an Individualistic Focus. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 87 (apr 2023), 34 pages. https://doi.org/10.1145/3579520

[2] Annalena Bea Aicher, Daniel Kornmüller, Wolfgang Minker, and Stefan Ultes. 2023. Self-imposed Filter Bubble Model for Argumentative Dialogues. *Proceedings of the 5th International Conference on Conversational User Interfaces* (2023), Article 23. Publisher: Association for Computing Machinery.

[3] Jennifer Allen, Cameron Martel, and David G Rand. 2022. Birds of a Feather Don't Fact-Check Each Other: Partisanship and the Evaluation of News in Twitter's Birdwatch Crowdsourced Fact-Checking Program. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022). https://doi.org/10.1145/3491102.3502040 Publisher: Association for Computing Machinery.

[4] Deemah Alqahtani, Caroline Jay, and Markel Vigo. 2020. The Role of Uncertainty as a Facilitator to Reflection in Self-Tracking. *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (2020), 1807–1818. https://doi.org/10.1145/3357236.3395448 Publisher: Association for Computing Machinery.

[5] Leif Azzopardi. 2021. Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval. *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (2021), 27–37. https://doi.org/10.1145/3406522.3446023 Publisher: Association for Computing Machinery.

[6] Leif Azzopardi and Jiqun Liu. 2024. Search under Uncertainty: Cognitive Biases and Heuristics - Tutorial on Modeling Search Interaction using Behavioral Economics. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval* (Sheffield, United Kingdom) *(CHIIR '24)*. Association for Computing Machinery, New York, NY, USA, 427–430. https://doi.org/10.1145/3627508.3638297

[7] Franziska Babel, Philipp Hock, Katie Winkle, Ilaria Torre, and Tom Ziemke. 2024. The Human Behind the Robot: Rethinking the Low Social Status of Service Robots. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '24)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3610978.3640763

[8] Anne Kathrine Petersen Bach, Trine Munch Nørgaard, Jens Christian Brok, and Niels van Berkel. 2023. "If I Had All the Time in the World": Ophthalmologists' Perceptions of Anchoring Bias Mitigation in Clinical AI Support. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023). https://doi.org/10.1145/3544548.3581513 Publisher: Association for Computing Machinery.

[9] Aruna D. Balakrishnan, Susan R. Fussell, Sara Kiesler, and Aniket Kittur. 2010. Pitfalls of Information Access with Visualizations in Remote Collaborative Analysis. *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work* (2010), 411–420. https://doi.org/10.1145/1718918.1718988 Publisher: Association for Computing Machinery.

[10] Oswald Barral, Gabor Aranyi, Sid Kouider, Alan Lindsay, Hielke Prins, Imtiaj Ahmed, Giulio Jacucci, Paolo Negri, Luciano Gamberini, David Pizzi, and Marc Cavazza. 2014. Covert Persuasive Technologies: Bringing Subliminal Cues to Human-Computer Interaction. In *Persuasive Technology*, Anna Spagnolli, Luca Chittaro, and Luciano Gamberini (Eds.). Springer International Publishing, Cham, 1–12.

[11] Eric P. S. Baumer, Jaime Snyder, and Geri K. Gay. 2018. Interpretive Impacts of Text Visualization: Mitigating Political Framing Effects. *ACM Trans. Comput.-Hum. Interact.* 25, 4 (2018). https://doi.org/10.1145/3214353

[12] Neil E. Beckwith and Donald R. Lehmann. 1975. The Importance of Halo Effects in Multi-Attribute Attitude Models. *Journal of Marketing Research* 12, 3 (1975), 265–275. http://www.jstor.org/stable/3151224

[13] H. Berghel. 2018. Malice Domestic: The Cambridge Analytica Dystopia. *Computer* 51, 05 (may 2018), 84–89. https://doi.org/10.1109/MC.2018.2381135

[14] Astrid Bertrand, Rafik Belloum, James R. Eagan, and Winston Maxwell. 2022. How Cognitive Biases Affect XAI-Assisted Decision-Making: A Systematic Review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 78–91. https://doi.org/10.1145/3514094.3534164

[15] Md Momen Bhuiyan, Michael Horning, Sang Won Lee, and Tanushree Mitra. 2021. NudgeCred: Supporting News Credibility Assessment on Social Media Through Nudges. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (2021). https://doi.org/10.1145/3479571

[16] Markus Bink, Sebastian Schwarz, Tim Draws, and David Elsweiler. 2023. Investigating the Influence of Featured Snippets on User Attitudes. *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval* (2023), 211–220. https://doi.org/10.1145/3576840.3578323 Publisher: Association for Computing Machinery.

[17] Nattapat Boonprakong, Xiuge Chen, Catherine Davey, Benjamin Tag, and Tilman Dingler. 2023. Bias-Aware Systems: Exploring Indicators for the Occurrences of Cognitive Biases When Facing Different Opinions. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023). https://doi.org/10.1145/3544548.3580917 Publisher: Association for Computing Machinery.

[18] Nattapat Boonprakong, Gaole He, Ujwal Gadiraju, Niels Van Berkel, Danding Wang, Si Chen, Jiqun Liu, Benjamin Tag, Jorge Goncalves, and Tilman Dingler. 2023. Workshop on Understanding and Mitigating Cognitive Biases in Human-AI Collaboration. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (Minneapolis, MN, USA) *(CSCW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 512–517. https://doi.org/10.1145/3584931.3611284

[19] Nattapat Boonprakong, Saumya Pareek, Benjamin Tag, Jorge Goncalves, and Tilman Dingler. 2025. Assessing Susceptibility Factors of Confirmation Bias in News Feed Reading. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3706598.3713873

[20] Nattapat Boonprakong, Benjamin Tag, and Tilman Dingler. 2023. Designing Technologies to Support Critical Thinking in an Age of Misinformation. *IEEE Pervasive Computing* (2023), 1–10. https://doi.org/10.1109/MPRV.2023.3275514

[21] Luc Bovens. 2009. *The Ethics of Nudge*. Springer Netherlands, Dordrecht, 207–219. https://doi.org/10.1007/978-90-481-2593-7_10

[22] Andreas Bulling and Thorsten O. Zander. 2014. Cognition-Aware Computing. *IEEE Pervasive Computing* 13, 3 (2014), 80–83. https://doi.org/10.1109/mprv.2014.42

[23] Pavlo Burda, Luca Allodi, and Nicola Zannone. 2024. Cognition in Social Engineering Empirical Research: A Systematic Literature Review. *ACM Trans. Comput.-Hum. Interact.* 31, 2 (2024). https://doi.org/10.1145/3635149

[24] Sahara Byrne and Philip Solomon Hart. 2009. The Boomerang Effect A Synthesis of Findings and a Preliminary Theoretical Framework. *Annals of the International Communication Association* 33, 1 (2009), 3–37. https://doi.org/10.1080/23808985.2009.11679083 arXiv:https://doi.org/10.1080/23808985.2009.11679083

[25] Federico Cabitza, Andrea Campagner, Riccardo Angius, Chiara Natali, and Carlo Reverberi. 2023. AI Shall Have No Dominion: On How to Measure Technology Dominance in AI-Supported Human Decision-Making. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023). https://doi.org/10.1145/3544548.3581095 Publisher: Association for Computing Machinery.

[26] Ángel Alexander Cabrera, Marco Tulio Ribeiro, Bongshin Lee, Robert Deline, Adam Perer, and Steven M. Drucker. 2023. What Did My AI Learn? How Data Scientists Make Sense of Model Behavior. *ACM Trans. Comput.-Hum. Interact.* 30, 1 (2023). https://doi.org/10.1145/3542921

[27] Shiye Cao, Anqi Liu, and Chien-Ming Huang. 2024. Designing for Appropriate Reliance: The Roles of AI Uncertainty Presentation, Initial User Decision, and User Demographics in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1 (2024). https://doi.org/10.1145/3637318

[28] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), 1–15. https://doi.org/10.1145/3290605.3300733 Publisher: Association for Computing Machinery.

[29] Ana Caraban, Loukas Konstantinou, and Evangelos Karapanos. 2020. The Nudge Deck: A Design Support Tool for Technology-Mediated Nudging. *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (2020), 395–406. https://doi.org/10.1145/3357236.3395485 Publisher: Association for Computing Machinery.

[30] John M. Carroll and Judith Reitman Olson. 1988. Mental Models in Human-Computer Interaction. In *Handbook of Human-Computer Interaction*, MARTIN HELANDER (Ed.). North-Holland, Amsterdam, 45–65. https://doi.org/10.1016/B978-0-444-70536-5.50007-5

[31] Serena Chen, Kimberly Duckworth, and Shelly Chaiken. 1999. Motivated Heuristic and Systematic Processing. *Psychological Inquiry* 10, 1 (1999), 44–49. http://www.jstor.org/stable/1449522

[32] Xiaogang Chen, Libo Su, and Darrell Carpenter. 2020. Impacts of Situational Factors on Consumers' Adoption of Mobile Payment Services: A Decision-Biases Perspective. *International Journal of Human–Computer Interaction* 36, 11 (2020), 1085–1093. https://doi.org/10.1080/10447318.2020.1722400

[33] Xiaojiao Chen, Xiaoteng Tang, Ying Zhao, Tengyu Huang, Ran Qian, Jiayi Zhang, Wei Chen, and Xiaosong Wang. 2024. Evaluating Visual Consistency of Icon Usage in Across-Devices. *International Journal of Human–Computer Interaction* 40, 9 (2024), 2415–2431. https://doi.org/10.1080/10447318.2022.2162275 Publisher: Taylor & Francis.

[34] Fu-Yin Cherng, Jingchao Fang, Yinhao Jiang, Xin Chen, Taejun Choi, and Hao-Chuan Wang. 2022. Understanding Social Influence in Collective Product Ratings Using Behavioral and Cognitive Metrics. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022). https://doi.org/10.1145/3491102.3517726 Publisher: Association for Computing Machinery.

[35] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2023. Are Two Heads Better Than One in AI-Assisted Decision Making? Comparing the Behavior and Performance of Groups and Individuals in Human-AI Collaborative Recidivism Risk Assessment. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023). https://doi.org/10.1145/3544548.3581015 Publisher: Association for Computing Machinery.

[36] Isaac Cho, Ryan Wesslen, Alireza Karduni, Sashank Santhanam, Samira Shaikh, and Wenwen Dou. 2017. The Anchoring Effect in Decision-Making with Visual Analytics. *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2017), 116–126. https://doi.org/10.1109/VAST.2017.8585665

[37] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. *26th International Conference on Intelligent User Interfaces* (2021), 307–317. https://doi.org/10.1145/3397481.3450644 Publisher: Association for Computing Machinery.

[38] Mark Colley, Jan Henry Belz, and Enrico Rukzio. 2021. Investigating the Effects of Feedback Communication of Autonomous Vehicles. *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (2021), 263–273. https://doi.org/10.1145/3409118.3475133 Publisher: Association for Computing Machinery.

[39] Thomas Constant and Guillaume Levieux. 2019. Dynamic Difficulty Adjustment Impact on Players' Confidence. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), 1–12. https://doi.org/10.1145/3290605.3300693 Publisher: Association for Computing Machinery.

[40] John Cook, Stephan Lewandowsky, and Ullrich K. H. Ecker. 2017. Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLOS ONE* 12, 5 (05 2017), 1–21.

https://doi.org/10.1371/journal.pone.0175799

[41] Florence L. Denmark and Deborah Williams. 2014. *Gender Bias, Overview.* Springer New York, New York, NY, 761–762. https://doi.org/10.1007/978-1-4614-5583-7_430

[42] Evanthia Dimara, Gilles Bailly, Anastasia Bezerianos, and Steven Franconeri. 2019. Mitigating the Attraction Effect with Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 850–860. https://doi.org/10.1109/TVCG.2018.2865233

[43] Evanthia Dimara, Anastasia Bezerianos, and Pierre Dragicevic. 2017. The Attraction Effect in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 471–480. https://doi.org/10.1109/TVCG.2016.2598594

[44] Evanthia Dimara, Steven Franconeri, Catherine Plaisant, Anastasia Bezerianos, and Pierre Dragicevic. 2020. A Task-Based Taxonomy of Cognitive Biases for Information Visualization. *IEEE Transactions on Visualization and Computer Graphics* 26, 2 (2020), 1413–1432. https://doi.org/10.1109/TVCG.2018.2872577

[45] Tilman Dingler, Benjamin Tag, David A. Eccles, Niels van Berkel, and Vassilis Kostakos. 2022. Method for Appropriating the Brief Implicit Association Test to Elicit Biases in Users. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022). https://doi.org/10.1145/3491102.3517570 Publisher: Association for Computing Machinery.

[46] Tilman Dingler, Benjamin Tag, Evangelos Karapanos, Koichi Kise, and Andreas Dengel. 2020. Workshop on Detection and Design for Cognitive Biases in People and Computing Systems. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3334480.3375159

[47] Tilman Dingler, Benjamin Tag, Philipp Lorenz-Spreen, Andrew W. Vargo, Simon Knight, and Stephan Lewandowsky. 2021. Workshop on Technologies to Support Critical Thinking in an Age of Misinformation. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–5. https://doi.org/10.1145/3411763.3441350

[48] Hongming Dong, Shoufeng Ma, Shuai Ling, Geng Li, Shuxian Xu, and Bo Song. 2023. An Empirical Investigation on the Acceptance of Autonomous Vehicles: Perspective of Drivers' Self–AV Bias. *International Journal of Human–Computer Interaction* 0, 0 (2023), 1–13. https://doi.org/10.1080/10447318.2023.2186000

[49] Tim Donkers and Jürgen Ziegler. 2021. The Dual Echo Chamber: Modeling Social Media Polarization for Interventional Recommending. *Proceedings of the 15th ACM Conference on Recommender Systems* (2021), 12–22. https://doi.org/10.1145/3460231.3474261 Publisher: Association for Computing Machinery.

[50] Tim Draws, Oana Inel, Nava Tintarev, Christian Baden, and Benjamin Timmermans. 2022. Comprehensive Viewpoint Representations for a Deeper Understanding of User Interactions With Debated Topics. *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval* (2022), 135–145. https://doi.org/10.1145/3498366.3505812 Publisher: Association for Computing Machinery.

[51] Tim Draws, Karthikeyan Natesan Ramamurthy, Ioana Baldini, Amit Dhurandhar, Inkit Padhi, Benjamin Timmermans, and Nava Tintarev. 2023. Explainable Cross-Topic Stance Detection for Search Results. *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval* (2023), 221–235. https://doi.org/10.1145/3576840.3578296 Publisher: Association for Computing Machinery.

[52] Jessica Maria Echterhoff, Matin Yarmand, and Julian McAuley. 2022. AI-Moderated Decision-Making: Capturing and Balancing Anchoring Bias in Sequential Decision Tasks. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022). https://doi.org/10.1145/3491102.3517443 Publisher: Association for Computing Machinery.

[53] Markus I. Eronen and Laura F. Bringmann. 2021. The Theory Crisis in Psychology: How to Move Forward. *Perspectives on psychological science : a journal of the Association for Psychological Science* 16, 4 (July 2021), 779–788. https://doi.org/10.1177/1745691620970586 Place: United States.

[54] Jonathon St BT Evans and David E Over. 1996. *Rationality and Reasoning.* Psychology Press.

[55] Jonathan St. B. T. Evans, Simon J. Handley, Helen Neilens, and David Over. 2010. The influence of cognitive ability and instructional set on causal conditional inference. *Quarterly Journal of Experimental Psychology* 63, 5 (2010), 892–909. https://doi.org/10.1080/17470210903111821 arXiv:https://doi.org/10.1080/17470210903111821 PMID: 19728225.

[56] Jonathan St. B. T. Evans and Keith E. Stanovich. 2013. Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science* 8, 3 (2013), 223–241. https://doi.org/10.1177/1745691612460685 arXiv:https://doi.org/10.1177/1745691612460685 PMID: 26172965.

[57] Lauren Fell, Andrew Gibson, Peter Bruza, and Pamela Hoyte. 2020. Human Information Interaction and the Cognitive Predicting Theory of Trust. *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval* (2020), 145–152. https://doi.org/10.1145/3343413.3377981 Publisher: Association for Computing Machinery.

[58] Cognitive Bias Foundation. 2024. Bias Cheat Sheet. http://bias.transhumanity.net/bias-cheat-sheet/. Accessed: 2024-09-13.

[59] Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel, and Stefan Dietze. 2017. Using Worker Self-Assessments for Competence-Based Pre-Selection in Crowdsourcing Microtasks. *ACM Trans. Comput.-Hum. Interact.* 24, 4 (2017). https://doi.org/10.1145/3119930

[60] Christine Geeng, Savanna Yee, and Franziska Roesner. 2020. Fake News on Facebook and Twitter: Investigating How People (Don't) Investigate. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), 1–14. Place: New York, NY, USA Publisher: Association for Computing Machinery.

[61] Amira Ghenai, Mark D. Smucker, and Charles L.A. Clarke. 2020. A Think-Aloud Study to Understand Factors Affecting Online Health Search. *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval* (2020), 273–282. https://doi.org/10.1145/3343413.3377961 Publisher: Association for Computing Machinery.

[62] Gerd Gigerenzer. 2004. *Fast and Frugal Heuristics: The Tools of Bounded Rationality.* John Wiley & Sons, Ltd, Chapter 4, 62–88. https://doi.org/10.1002/9780470752937.ch4 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470752937.ch4

[63] Gerd Gigerenzer. 2008. Why Heuristics Work. *Perspectives on Psychological Science* 3, 1 (2008), 20–29. https://doi.org/10.1111/j.1745-6916.2008.00058.x arXiv:https://doi.org/10.1111/j.1745-6916.2008.00058.x PMID: 26158666.

[64] Thomas Gilovich, Dale Griffin, and Daniel Kahneman. 2002. *Heuristics and Biases: The Psychology of Intuitive Judgment.* Cambridge University Press.

[65] Klaus Goffart, Michael Schermann, Christopher Kohl, Jörg Preißinger, and Helmut Krcmar. 2016. Using the Default Option Bias to Influence Decision Making While Driving. *International Journal of Human–Computer Interaction* 32, 1 (2016), 39–50. https://doi.org/10.1080/10447318.2015.1085747

[66] Navita Goyal, Connor Baumler, Tin Nguyen, and Hal Daumé III. 2024. The Impact of Explanations on Fairness in Human-AI Decision-Making: Protected vs Proxy Features. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 155–180. https://doi.org/10.1145/3640543.3645210

[67] Eduardo Graells-Garrido, Mounia Lalmas, and Ricardo Baeza-Yates. 2016. Data Portraits and Intermediary Topics: Encouraging Exploration of Politically Diverse Profiles. *Proceedings of the 21st International Conference on Intelligent User Interfaces* (2016), 228–240. https://doi.org/10.1145/2856767.2856776 Publisher: Association for Computing Machinery.

[68] Sukeshini A. Grandhi, Linda Plotnick, and Starr Roxanne Hiltz. 2019. Do I Stay or Do I Go? Motivations and Decision Making in Social Media Non-Use and Reversion. *Proc. ACM Hum.-Comput. Interact.* 3, GROUP (2019). https://doi.org/10.1145/3361116

[69] Aditya Gulati, Miguel Angel Lozano, Bruno Lepri, and Nuria Oliver. 2023. BIASeD: Bringing Irrationality into Automated System Design. CEUR. http://hdl.handle.net/10045/132001

[70] Xunhua Guo, Lingli Wang, Mingyue Zhang, and Guoqing Chen. 2023. First Things First? Order Effects in Online Product Recommender Systems. *ACM Trans. Comput.-Hum. Interact.* 30, 1 (2023). https://doi.org/10.1145/3557886

[71] Taehyun Ha and Sangyeon Kim. 2024. Improving Trust in AI with Mitigating Confirmation Bias: Effects of Explanation Type and Debiasing Strategy for Decision-Making with Explainable AI. *International Journal of Human–Computer Interaction* (2024), 1–12. https://doi.org/10.1080/10447318.2023.2285640 Publisher: Taylor & Francis.

[72] Hana Habib, Megan Li, Ellie Young, and Lorrie Cranor. 2022. "Okay, Whatever": An Evaluation of Cookie Consent Interfaces. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022). https://doi.org/10.1145/3491102.3501985 Publisher: Association for Computing Machinery.

[73] Sophia Hadash, Martijn C. Willemsen, Chris Snijders, and Wijnand A. IJsselsteijn. 2022. Improving Understandability of Feature Contributions in Model-Agnostic Explainable AI Tools. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022). https://doi.org/10.1145/3491102.3517650 Publisher: Association for Computing Machinery.

[74] Martin S. Hagger. 2014. Avoiding the "déjà-variable" phenomenon: social psychology needs more guides to constructs. *Frontiers in psychology* 5 (2014), 52. https://doi.org/10.3389/fpsyg.2014.00052 Place: Switzerland.

[75] Ehsan-Ul Haq, Yang K. Lu, and Pan Hui. 2022. It's All Relative! A Method to Counter Human Bias in Crowdsourced Stance Detection of News Articles. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2 (2022). https://doi.org/10.1145/3555636

[76] MD Romael Haque, Devansh Saxena, Katy Weathington, Joseph Chudzik, and Shion Guha. 2024. Are We Asking the Right Questions?: Designing for Community Stakeholders' Interactions with AI in Policing. *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024). https://doi.org/10.1145/3613904.3642738 Publisher: Association for Computing Machinery.

[77] Christopher G. Harris. 2019. Detecting Cognitive Bias in a Relevance Assessment Task Using an Eye Tracker. *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications* (2019). https://doi.org/10.1145/3314111.3319824 Publisher: Association for Computing Machinery.

[78] Martie G. Haselton, Gregory A. Bryant, Andreas Wilke, David A. Frederick, Andrew Galperin, Willem E. Frankenhuis, and Tyler Moore. 2009. Adaptive Rationality: An Evolutionary Perspective on Cognitive Bias. *Social Cognition* 27, 5 (2009), 733–763. https://doi.org/10.1521/soco.2009.27.5.733 arXiv:https://doi.org/10.1521/soco.2009.27.5.733

[79] Martie G Haselton, Daniel Nettle, and Damian R Murray. 2015. The evolution of cognitive bias. *The handbook of evolutionary psychology* (2015), 1–20.

[80] Yugo Hayashi, Kosuke Wakabayashi, and Yuki Nishida. 2023. How Sequential Suggestions from a Robot and Human Jury Influence Decision Making: A Large Scale Investigation Using a Court Sentencing Judgment Task. *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (2023), 338–341. https://doi.org/10.1145/3568294.3580101 Publisher: Association for Computing Machinery.

[81] Gaole He, Stefan Buijsman, and Ujwal Gadiraju. 2023. How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 276 (Oct. 2023), 29 pages. https://doi.org/10.1145/3610067

[82] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023). https://doi.org/10.1145/3544548.3581025 Publisher: Association for Computing Machinery.

[83] Eric B. Hekler, Predrag Klasnja, Jon E. Froehlich, and Matthew P. Buman. 2013. Mind the Theoretical Gap: Interpreting, Using, and Developing Behavioral Theory in HCI Research. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2013), 3307–3316. https://doi.org/10.1145/2470654.2466452 Publisher: Association for Computing Machinery.

[84] Sarita Herse, Jonathan Vitale, and Mary-Anne Williams. 2023. Using Agent Features to Influence User Trust, Decision Making and Task Outcome during Human-Agent Collaboration. *International Journal of Human–Computer Interaction* 39, 9 (2023), 1740–1761. https://doi.org/10.1080/10447318.2022.2150691 Publisher: Taylor & Francis.

[85] Ralph Hertwig and Till Grüne-Yanoff. 2017. Nudging and Boosting: Steering or Empowering Good Decisions. *Perspectives on Psychological Science* 12, 6 (2017), 973–986. https://doi.org/10.1177/1745691617702496 arXiv:https://doi.org/10.1177/1745691617702496 PMID: 28792862.

[86] Martin Hilbert. 2012. Toward a synthesis of cognitive biases: how noisy information processing can bias human decision making. *Psychological bulletin* 138, 2 (2012), 211.

[87] Mohamad Hjeij and Arnis Vilks. 2023. A brief history of heuristics: how did research on heuristics evolve? *Humanities and Social Sciences Communications* 10, 1 (Feb. 2023), 64. https://doi.org/10.1057/s41599-023-01542-z

[88] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), 1–12. https://doi.org/10.1145/3290605.3300637 Publisher: Association for Computing Machinery.

[89] Jessica Hullman, Eytan Adar, and Priti Shah. 2011. The impact of social information on visual judgments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 1461–1470. https://doi.org/10.1145/1978942.1979157

[90] David G. Jansson and Steven M. Smith. 1991. Design fixation. *Design Studies* 12, 1 (1991), 3–11. https://doi.org/10.1016/0142-694X(91)90003-F

[91] Gabbrielle M. Johnson. 2021. Algorithmic bias: on the implicit biases of social technology. *Synthese* 198, 10 (Oct. 2021), 9941–9961. https://doi.org/10.1007/s11229-020-02696-y

[92] Christine Jolls and Cass R. Sunstein. 2006. Debiasing through Law. *The Journal of Legal Studies* 35, 1 (2006), 199–242. https://doi.org/10.1086/500096 arXiv:https://doi.org/10.1086/500096

[93] Eva Jonas, Stefan Schulz-Hardt, Dieter Frey, and Norman Thelen. 2001. Confirmation bias in sequential information search after preliminary decisions: an expansion of dissonance theoretical research on selective exposure to information. *Journal of personality and social psychology* 80, 4 (2001), 557.

[94] Emma B. Jones and Louise Sharpe. 2017. Cognitive bias modification: A review of meta-analyses. *Journal of Affective Disorders* 223 (2017), 175–183. https://doi.org/10.1016/j.jad.2017.07.034

[95] Daniel Kahneman. 2011. *Thinking, Fast and Slow.* Macmillan.

[96] Daniel Kahneman, Olivier Sibony, and Cass R Sunstein. 2021. *Noise: A flaw in human judgment.* Hachette UK.

[97] Daniel Kahneman and Amos Tversky. 1972. Subjective probability: A judgment of representativeness. *Cognitive psychology* 3, 3 (1972), 430–454.

[98] Daniel Kahneman and Amos Tversky. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47, 2 (1979), 263–291. http://www.jstor.org/stable/1914185

[99] Naomi Kakoschke, Rowan Page, Barbora de Courten, Antonio Verdejo-Garcia, and Jon McCormack. 2021. Brain training with the body in mind: Towards gamified approach-avoidance training using virtual reality. *International Journal of Human-Computer Studies* 151 (2021), 102626. https://doi.org/10.1016/j.ijhcs.2021.102626

[100] Alireza Karduni, Isaac Cho, Ryan Wesslen, Sashank Santhanam, Svitlana Volkova, Dustin L Arendt, Samira Shaikh, and Wenwen Dou. 2019. Vulnerable to Misinformation? Verifi! *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019), 312–323. https://doi.org/10.1145/3301275.3302320 Publisher: Association for Computing Machinery.

[101] Harmanpreet Kaur, Matthew R. Conrad, Davis Rule, Cliff Lampe, and Eric Gilbert. 2024. Interpretability Gone Bad: The Role of Bounded Rationality in How Practitioners Understand Machine Learning. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 77 (apr 2024), 34 pages. https://doi.org/10.1145/3637354

[102] Jieun Kim, Hokyoung Ryu, and Hyeonah Kim. 2013. To be biased or not to be: choosing between design fixation and design intentionality. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (Paris, France) *(CHI EA '13)*. Association for Computing Machinery, New York, NY, USA, 349–354. https://doi.org/10.1145/2468356.2468418

[103] Taenyun Kim and Hayeon Song. 2023. Communicating the Limitations of AI: The Effect of Message Framing and Ownership on Trust in Artificial Intelligence. *International Journal of Human–Computer Interaction* 39, 4 (2023), 790–800. https://doi.org/10.1080/10447318.2022.2049134

[104] Tomáš Kliegr, Štěpán Bahník, and Johannes Fürnkranz. 2021. A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence* 295 (2021), 103458. https://doi.org/10.1016/j.artint.2021.103458

[105] Sara Klüber, Franzisca Maas, David Schraudt, Gina Hermann, Oliver Happel, and Tobias Grundgeiger. 2020. Experience Matters: Design and Evaluation of an Anesthesia Support Tool Guided by User Experience Theory. *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (2020), 1523–1535. https://doi.org/10.1145/3357236.3395552 Publisher: Association for Computing Machinery.

[106] Silvia Knobloch-Westerwick, Benjamin K. Johnson, Nathaniel A. Silver, and Axel Westerwick. 2015. Science Exemplars in the Eye of the Beholder: How Exposure to Online Science Information Affects Attitudes. *Science Communication* 37, 5 (2015), 575–601. https://doi.org/10.1177/1075547015596367 arXiv:https://doi.org/10.1177/1075547015596367

[107] Ha-Kyung Kong, Zhicheng Liu, and Karrie Karahalios. 2018. Frames and Slants in Titles of Visualizations on Controversial Topics. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), 1–12. https://doi.org/10.1145/3173574.3174012 Publisher: Association for Computing Machinery.

[108] Ha-Kyung Kong, Zhicheng Liu, and Karrie Karahalios. 2019. Trust and Recall of Information across Varying Degrees of Title-Visualization Misalignment. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), 1–13. https://doi.org/10.1145/3290605.3300576 Publisher: Association for Computing Machinery.

[109] Loukas Konstantinou, Dionysis Panos, and Evangelos Karapanos. 2024. Exploring the Design of Technology-Mediated Nudges for Online Misinformation. *International Journal of Human–Computer Interaction* (2024), 1–28. https://doi.org/10.1080/10447318.2023.2301265 Publisher: Taylor & Francis.

[110] Thomas Kosch, Robin Welsch, Lewis Chuang, and Albrecht Schmidt. 2023. The Placebo Effect of Artificial Intelligence in Human–Computer Interaction. *ACM Trans. Comput.-Hum. Interact.* 29, 6 (2023). https://doi.org/10.1145/3529225

[111] Morgane Koval and Yvonne Jansen. 2022. Do You See What You Mean? Using Predictive Visualizations to Reduce Optimism in Duration Estimates. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022). https://doi.org/10.1145/3491102.3502010 Publisher: Association for Computing Machinery.

[112] Anastasia Kozyreva, Stephan Lewandowsky, and Ralph Hertwig. 2020. Citizens Versus the Internet: Confronting Digital Challenges With Cognitive Tools. *Psychological Science in the Public Interest* 21, 3 (2020), 103–156. https://doi.org/10.1177/1529100620946707 arXiv:https://doi.org/10.1177/1529100620946707 PMID: 33325331.

[113] Justin Kruger. 1999. Lake Wobegon be gone! The" below-average effect" and the egocentric nature of comparative ability judgments. *Journal of personality and social psychology* 77, 2 (1999), 221.

[114] Minae Kwon, Erdem Biyik, Aditi Talati, Karan Bhasin, Dylan P. Losey, and Dorsa Sadigh. 2020. When Humans Aren't Optimal: Robots That Collaborate with Risk-Aware Humans. *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (2020), 43–52. https://doi.org/10.1145/3319502.3374832 Publisher: Association for Computing Machinery.

[115] Mitra Lashkari and Jinghui Cheng. 2023. "Finding the Magic Sauce": Exploring Perspectives of Recruiters and Job Seekers on Recruitment Bias and Automated Tools. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023). https://doi.org/10.1145/3544548.3581548 Publisher: Association for Computing Machinery.

[116] Nguyen-Thinh Le and Laura Wartschinski. 2018. A Cognitive Assistant for improving human reasoning skills. *International Journal of Human-Computer Studies* 117 (2018), 45–54. https://doi.org/10.1016/j.ijhcs.2018.02.005

[117] Kwan Min Lee, Younbo Jung, Jaywoo Kim, and Sang Ryong Kim. 2006. Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *International Journal of Human-Computer Studies* 64, 10 (2006), 962–973. https://doi.org/10.1016/j.ijhcs.2006.05.002

[118] Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. 2011. Mining Behavioral Economics to Design Persuasive Technology for Healthy Choices. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2011), 325–334. https://doi.org/10.1145/1978942.1978989 Publisher: Association for Computing Machinery.

[119] Stephan Lewandowsky, Ullrich K.H. Ecker, and John Cook. 2017. Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era. *Journal of Applied Research in Memory and Cognition* 6, 4 (2017), 353–369. https://doi.org/10.1016/j.jarmac.2017.07.008

[120] Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest* 13, 3 (2012), 106–131. https://doi.org/10.1177/1529100612451018 arXiv:https://doi.org/10.1177/1529100612451018 PMID: 26173286.

[121] Tony W Li, Arshia Arya, and Haojian Jin. 2024. Redesigning Privacy with User Feedback: The Case of Zoom Attendee Attention Tracking. *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024). https://doi.org/10.1145/3613904.3642594 Publisher: Association for Computing Machinery.

[122] Q. Vera Liao and Wai-Tat Fu. 2014. Can you hear me now? mitigating the echo chamber effect by source position indicators. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Baltimore, Maryland, USA) *(CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 184–196. https://doi.org/10.1145/2531602.2531711

[123] Q. Vera Liao, Wai-Tat Fu, and Sri Shilpa Mamidi. 2015. It Is All About Perspective: An Exploration of Mitigating Selective Exposure with Aspect Indicators. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015), 1439–1448. https://doi.org/10.1145/2702123.2702570 Publisher: Association for Computing Machinery.

[124] Scott O. Lilienfeld, Rachel Ammirati, and Kristin Landfield. 2009. Giving Debiasing Away: Can Psychological Research on Correcting Cognitive Errors Promote Human Welfare? *Perspectives on Psychological Science* 4, 4 (2009), 390–398. https://doi.org/10.1111/j.1745-6924.2009.01144.x arXiv:https://doi.org/10.1111/j.1745-6924.2009.01144.x PMID: 26158987.

[125] Jiqun Liu. 2023. Toward A Two-Sided Fairness Framework in Search and Recommendation. *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval* (2023), 236–246. https://doi.org/10.1145/3576840.3578332 Publisher: Association for Computing Machinery.

[126] Qianyu Liu, Haoran Jiang, Zihao Pan, Qiushi Han, Zhenhui Peng, and Quan Li. 2024. BiasEye: A Bias-Aware Real-time Interactive Material Screening System for Impartial Candidate Assessment. *Proceedings of the 29th International Conference on Intelligent User Interfaces* (2024), 325–343. https://doi.org/10.1145/3640543.3645166 Publisher: Association for Computing Machinery.

[127] Meagan B. Loerakker, Jasmin Niess, Marit Bentvelzen, and Paweł W. Woźniak. 2024. Designing Data Visualisations for Self-Compassion in Personal Informatics. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 4 (2024), Article 169.

[128] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023). https://doi.org/10.1145/3544548.3581058 Publisher: Association for Computing Machinery.

[129] Shuai Ma, Xinru Wang, Ying Lei, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2024. "Are You Really Sure?" Understanding the Effects of Human Self-Confidence Calibration in AI-Assisted Decision Making. *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024). https://doi.org/10.1145/3613904.3642671 Publisher: Association for Computing Machinery.

[130] Peter G. Mahon and Roxanne L. Canosa. 2012. Prisoners and chickens: gaze locations indicate bounded rationality. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (Santa Barbara, California) *(ETRA '12)*. Association for Computing Machinery, New York, NY, USA, 401–404. https://doi.org/10.1145/2168556.2168647

[131] Prateek Mantri, Hariharan Subramonyam, Audrey L. Michal, and Cindy Xiong. 2023. How Do Viewers Synthesize Conflicting Information from Data Visualizations? *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2023), 1005–1015. https://doi.org/10.1109/TVCG.2022.3209467

[132] Arunesh Mathur, Gunes Acar, Michael J. Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (2019). https://doi.org/10.1145/3359183

[133] Arunesh Mathur, Mihir Kshirsagar, and Jonathan Mayer. 2021. What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021). https://doi.org/10.1145/3411764.3445610 Publisher: Association for Computing Machinery.

[134] Gonzalo Mendez, Luis Galárraga, and Katherine Chiluiza. 2021. Showing Academic Performance Predictions during Term Planning: Effects on Students' Decisions, Behaviors, and Preferences. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021). https://doi.org/10.1145/3411764.3445718 Publisher: Association for Computing Machinery.

[135] Ronald A. Metoyer, Tee Chuanromanee, Gina M. Girgis, Qiyu Zhi, and Eleanor C. Kinyon. 2020. Supporting Storytelling With Evidence in Holistic Review Processes: A Participatory Design Approach. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1 (2020). https://doi.org/10.1145/3392870

[136] Thomas Mildner, Albert Inkoom, Rainer Malaka, and Jasmin Niess. 2024. Hell is Paved with Good Intentions: The Intricate Relationship Between Cognitive Biases and Dark Patterns. arXiv:2405.07378 [cs.HC] https://arxiv.org/abs/2405.07378

[137] David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas Altman. 2009. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the PRISMA statement. *Br Med J* 8 (07 2009), 336–341. https://doi.org/10.1371/joumal.pmedl000097

[138] Zachary Munn, Micah D J Peters, Cindy Stern, Catalin Tufanaru, Alexa McArthur, and Edoardo Aromataris. 2018. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med. Res. Methodol.* 18, 1 (Nov. 2018), 143.

[139] Mohammad Naiseh, Dena Al-Thani, Nan Jiang, and Raian Ali. 2023. How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies* 169 (2023), 102941. https://doi.org/10.1016/j.ijhcs.2022.102941

[140] Arpit Narechania, Adam Coscia, Emily Wall, and Alex Endert. 2022. Lumos: Increasing Awareness of Analytic Behavior during Visual Data Analysis. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2022), 1009–1018. https://doi.org/10.1109/TVCG.2021.3114827

[141] Feng Ni, David Arnott, and Shijia Gao. 2019. The anchoring effect in business intelligence supported decision-making. *Journal of Decision Systems* 28, 2 (2019), 67–81. https://doi.org/10.1080/12460125.2019.1620573 arXiv:https://doi.org/10.1080/12460125.2019.1620573

[142] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175–220.

[143] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. *26th International Conference on Intelligent User Interfaces* (2021), 340–350. https://doi.org/10.1145/3397481.3450639 Publisher: Association for Computing Machinery.

[144] Alamir Novin and Eric Meyers. 2017. Making Sense of Conflicting Science Information: Exploring Bias in the Search Engine Result Page. *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (2017), 175–184. https://doi.org/10.1145/3020165.3020185 Publisher: Association for Computing Machinery.

[145] Alexander Nussbaumer, Katrien Verbert, Eva-Catherine Hillemann, Michael A. Bedek, and Dietrich Albert. 2016. A Framework for Cognitive Bias Detection and Feedback in a Visual Analytics Environment. *2016 European Intelligence and Security Informatics Conference (EISIC)* (2016), 148–151. https://doi.org/10.1109/EISIC.2016.038

[146] Tobias Nyström and Moyen M. Mustaquim. 2015. Managing Framing Effects in Persuasive Design for Sustainability. *Proceedings of the 19th International Academic Mindtrek Conference* (2015), 122–129. https://doi.org/10.1145/2818187.2818277 Publisher: Association for Computing Machinery.

[147] Aileen Oeberst and Roland Imhoff. 2023. Toward Parsimony in Bias Research: A Proposed Common Framework of Belief-Consistent Information Processing for a Set of Biases. *Perspectives on Psychological Science* 18, 6 (2023), 1464–1487. https://doi.org/10.1177/17456916221148147 arXiv:https://doi.org/10.1177/17456916221148147 PMID: 36930530.

[148] Tadashi Okoshi, Wataru Sasaki, and Jin Nakazawa. 2020. Behavification: bypassing human's attentional and cognitive systems for automated behavior change. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers* (Virtual Event, Mexico) *(UbiComp/ISWC '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 692–695. https://doi.org/10.1145/3410530.3414439

[149] Shaul Oreg and Mahmut Bayazit. 2009. Prone to Bias: Development of a Bias Taxonomy from an Individual Differences Perspective. *Review of General Psychology* 13, 3 (2009), 175–193. https://doi.org/10.1037/a0015656 arXiv:https://doi.org/10.1037/a0015656

[150] Steffi Paepcke and Leila Takayama. 2010. Judging a Bot by Its Cover: An Experiment on Expectation Setting for Personal Robots. *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction* (2010), 45–52. Publisher: IEEE Press.

[151] Marvin Pafla, Kate Larson, and Mark Hancock. 2024. Unraveling the Dilemma of AI Errors: Exploring the Effectiveness of Human and Machine Explanations for Large Language Models. *Proceedings of the CHI Conference on Human Factors in*

Computing Systems (2024). https://doi.org/10.1145/3613904.3642934 Publisher: Association for Computing Machinery.

[152] Charlie Pinder, Jo Vermeulen, Benjamin R. Cowan, and Russell Beale. 2018. Digital Behaviour Change Interventions to Break and Form Habits. *ACM Trans. Comput.-Hum. Interact.* 25, 3 (2018). https://doi.org/10.1145/3196830

[153] Suppanut Pothirattanachaikul, Takehiro Yamamoto, Yusuke Yamamoto, and Masatoshi Yoshikawa. 2020. Analyzing the Effects of "People Also Ask" on Search Behaviors and Beliefs. *Proceedings of the 31st ACM Conference on Hypertext and Social Media* (2020), 101–110. https://doi.org/10.1145/3372923.3404786 Publisher: Association for Computing Machinery.

[154] Marianne Procopio, Ab Mosca, Carlos Scheidegger, Eugene Wu, and Remco Chang. 2022. Impact of Cognitive Biases on Progressive Visualization. *IEEE Transactions on Visualization and Computer Graphics* 28, 9 (2022), 3093–3112. https://doi.org/10.1109/TVCG.2021.3051013

[155] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. 2022. Deciding Fast and Slow: The Role of Cognitive Biases in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1 (2022). https://doi.org/10.1145/3512930

[156] Karen Renaud and Verena Zimmermann. 2018. Ethical guidelines for nudging in information security & privacy. *International Journal of Human-Computer Studies* 120 (2018), 22–35. https://doi.org/10.1016/j.ijhcs.2018.05.011

[157] Eugenia Ha Rim Rho, Gloria Mark, and Melissa Mazmanian. 2018. Fostering Civil Discourse Online: Linguistic Behavior in Comments of #MeToo Articles across Political Perspectives. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (2018). https://doi.org/10.1145/3274416

[158] Alisa Rieger, Tim Draws, Mariët Theune, and Nava Tintarev. 2021. This Item Might Reinforce Your Opinion: Obfuscation and Labeling of Search Results to Mitigate Confirmation Bias. *Proceedings of the 32nd ACM Conference on Hypertext and Social Media* (2021), 189–199. https://doi.org/10.1145/3465336.3475101 Publisher: Association for Computing Machinery.

[159] Alisa Rieger, Qurat-Ul-Ain Shaheen, Carles Sierra, Mariet Theune, and Nava Tintarev. 2022. Towards Healthy Engagement with Online Debates: An Investigation of Debate Summaries and Personalized Persuasive Suggestions. *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization* (2022), 192–199. https://doi.org/10.1145/3511047.3537692 Publisher: Association for Computing Machinery.

[160] Lee Ross. 1977. The Intuitive Psychologist And His Shortcomings: Distortions in the Attribution Process. Advances in Experimental Social Psychology, Vol. 10. Academic Press, 173–220. https://doi.org/10.1016/S0065-2601(08)60357-3

[161] Ulrich Schimmack. 2020. A Meta-Scientific Perspective on "Thinking: Fast and Slow. https://replicationindex.com/2020/12/30/a-meta-scientific-perspective-on-thinking-fast-and-slow/. Accessed: 2024-11-08.

[162] Christina Schwind, Jürgen Buder, and Friedrich W. Hesse. 2011. I Will Do It, but i Don't like It: User Reactions to Preference-Inconsistent Recommendations. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2011), 349–352. https://doi.org/10.1145/1978942.1978992 Publisher: Association for Computing Machinery.

[163] Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024). https://doi.org/10.1145/3613904.3642459 Publisher: Association for Computing Machinery.

[164] Li Shi, Nilavra Bhattacharya, Anubrata Das, and Jacek Gwizdka. 2023. True or False? Cognitive Load When Reading COVID-19 News Headlines: An Eye-Tracking Study. *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval* (2023), 107–116. https://doi.org/10.1145/3576840.3578290 Publisher: Association for Computing Machinery.

[165] Wenxuan Wendy Shi, Sneha R. Krishna Kumaran, Hari Sundaram, and Brian P. Bailey. 2023. The Value of Activity Traces in Peer Evaluations: An Experimental Study. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1 (2023). https://doi.org/10.1145/3579627

[166] Herbert A Simon. 1957. A behavioral model of rational choice. *Models of man, social and rational: Mathematical essays on rational human behavior in a social setting* (1957), 241–260.

[167] Saniat Javid Sohrawardi, Y. Kelly Wu, Andrea Hickerson, and Matthew Wright. 2024. Dungeons & Deepfakes: Using scenario-based role-play to study journalists' behavior towards using AI-based verification tools for video content. *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024). https://doi.org/10.1145/3613904.3641973 Publisher: Association for Computing Machinery.

[168] Jack B. Soll, Katherine L. Milkman, and John W. Payne. 2015. *A User's Guide to Debiasing.* John Wiley & Sons, Ltd, Chapter 33, 924–951. https://doi.org/10.1002/9781118468333.ch33 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118468333.ch33

[169] Jacob Solomon. 2014. Customization Bias in Decision Support Systems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2014), 3065–3074. https://doi.org/10.1145/2556288.2557211 Publisher: Association for Computing Machinery.

[170] Paul M. Spengler, Douglas C. Strohmer, David N. Dixon, and Victoria A. Shivy. 1995. A Scientist-Practitioner Model of Psychological Assessment: Implications for Training, Practice and Research. *The Counseling Psychologist* 23, 3 (1995), 506–534. https://doi.org/10.1177/0011000095233009 arXiv:https://doi.org/10.1177/0011000095233009

[171] Keith E Stanovich. 1999. *Who is Rational?: Studies of Individual Differences in Reasoning.* Psychology Press.

[172] Keith E. Stanovich and Richard F. West. 2000. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences* 23, 5 (2000), 645–665. https://doi.org/10.1017/S0140525X00003435

[173] Poorna Talkad Sukumar, Ronald Metoyer, and Shuai He. 2018. Making a Pecan Pie: Understanding and Supporting The Holistic Review Process in Admissions. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (2018). https://doi.org/10.1145/3274438

[174] Lu Sun, Hengyuan Zhang, Enze Liu, Mingyang Liu, and Kristen Vaccaro. 2024. NewsGuesser: Using Curiosity to Reduce Selective Exposure. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1 (2024). https://doi.org/10.1145/3637376

[175] Cass R Sunstein. 2015. Nudging and choice architecture: Ethical considerations. *Yale Journal on Regulation, Forthcoming* (2015).

[176] Cass R. Sunstein and Richard H. Thaler. 2003. Libertarian Paternalism Is Not an Oxymoron. *The University of Chicago Law Review* 70, 4 (2003), 1159–1202. http://www.jstor.org/stable/1600573

[177] Pang Suwanaposee, Carl Gutwin, Zhe Chen, and Andy Cockburn. 2023. 'Specially For You' – Examining the Barnum Effect's Influence on the Perceived Quality of System Recommendations. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023). https://doi.org/10.1145/3544548.3580656 Publisher: Association for Computing Machinery.

[178] Siddharth Swaroop, Zana Buçinca, Krzysztof Z. Gajos, and Finale Doshi-Velez. 2024. Accuracy-Time Tradeoffs in AI-Assisted Decision Making under Time Pressure. *Proceedings of the 29th International Conference on Intelligent User Interfaces* (2024), 138–154. https://doi.org/10.1145/3640543.3645206 Publisher: Association for Computing Machinery.

[179] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, Textual or Hybrid: The Effect of User Expertise on Different Explanations. *26th International Conference on Intelligent User Interfaces* (2021), 109–119. https://doi.org/10.1145/3397481.3450662 Publisher: Association for Computing Machinery.

[180] Richard H. Thaler. 2018. Nudge, not sludge. *Science* 361, 6401 (2018), 431–431. https://doi.org/10.1126/science.aau9241 arXiv:https://www.science.org/doi/pdf/10.1126/science.aau9241

[181] R H Thaler and C R Sunstein. 2009. *Nudge: Improving Decisions About Health, Wealth, and Happiness.* Penguin Publishing Group.

[182] Georgios Theocharous, Jennifer Healey, Sridhar Mahadevan, and Michele Saad. 2019. Personalizing with Human Cognitive Biases. *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization* (2019), 13–17. https://doi.org/10.1145/3314183.3323453 Publisher: Association for Computing Machinery.

[183] Ian Thomas, Song Young Oh, and Danielle Albers Szafir. 2024. Assessing User Trust in Active Learning Systems: Insights from Query Policy and Uncertainty Visualization. *Proceedings of the 29th International Conference on Intelligent User Interfaces* (2024), 772–786. https://doi.org/10.1145/3640543.3645207 Publisher: Association for Computing Machinery.

[184] Paul Thomas, Gabriella Kazai, Ryen White, and Nick Craswell. 2022. The Crowd is Made of People: Observations from Large-Scale Crowd Labelling. *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval* (2022), 25–35. https://doi.org/10.1145/3498366.3505815 Publisher: Association for Computing Machinery.

[185] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. 2022. Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022). https://doi.org/10.1145/3491102.3517732 Publisher: Association for Computing Machinery.

[186] Natanael Bandeira Romão Tomé, Madison Klarkowski, Carl Gutwin, Cody Phillips, Regan L. Mandryk, and Andy Cockburn. 2020. Risking Treasure: Testing Loss Aversion in an Adventure Game. *Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (2020), 306–320. https://doi.org/10.1145/3410404.3414250 Publisher: Association for Computing Machinery.

[187] Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* 5, 2 (1973), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

[188] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131.

[189] Amos Tversky and Daniel Kahneman. 1988. Rational choice and the framing of decisions. *Decision making: Descriptive, normative, and prescriptive interactions* (1988), 167–192.

[190] Niels van Berkel, Maura Bellio, Mikael B. Skov, and Ann Blandford. 2023. Measurements, Algorithms, and Presentations of Reality: Framing Interactions with AI-Enabled Decision Support. *ACM Trans. Comput.-Hum. Interact.* 30, 2 (2023).

https://doi.org/10.1145/3571815

[191] Elizabeth S. Veinott, James Leonard, Elizabeth Lerner Papautsky, Brandon Perelman, Aleksandra Stankovic, Jared Lorince, Jared Hotaling, Travis Ross, Peter Todd, Edward Castronova, Jerome Busemeyer, Christoper Hale, Richard Catrambone, Elizabeth Whitaker, Olivia Fox, John Flach, and Robert R. Hoffman. 2013. The effect of camera perspective and session duration on training decision making in a serious video game. *2013 IEEE International Games Innovation Conference (IGIC)* (2013), 256–262. https://doi.org/10.1109/IGIC.2013.6659170

[192] Arnav Verma, Luiz Morais, Pierre Dragicevic, and Fanny Chevalier. 2023. Designing Resource Allocation Tools to Promote Fair Allocation: Do Visualization and Information Framing Matter? *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023). https://doi.org/10.1145/3544548.3580739 Publisher: Association for Computing Machinery.

[193] Julien Veytizou, David Bertolo, Charlotte Baraudon, Alexis Olry, and Stéphanie Fleck. 2018. Could a Tangible Interface Help a Child to Weigh His/Her Opinion on Usability? *Proceedings of the 30th Conference on l'Interaction Homme-Machine* (2018), 12–19. https://doi.org/10.1145/3286689.3286702 Publisher: Association for Computing Machinery.

[194] Peter B.M. Vranas. 2000. Gigerenzer's normative critique of Kahneman and Tversky. *Cognition* 76, 3 (2000), 179–193. https://doi.org/10.1016/S0010-0277(99)00084-0

[195] Samangi Wadinambiarachchi, Ryan M. Kelly, Saumya Pareek, Qiushi Zhou, and Eduardo Velloso. 2024. The Effects of Generative AI on Design Fixation and Divergent Thinking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 380, 18 pages. https://doi.org/10.1145/3613904.3642919

[196] Emily Wall, Arup Arcalgud, Kuhu Gupta, and Andrew Jo. 2019. A Markov Model of Users' Interactive Behavior in Scatterplots. *2019 IEEE Visualization Conference (VIS)* (2019), 81–85. https://doi.org/10.1109/VISUAL.2019.8933779

[197] Emily Wall, Leslie M. Blaha, Lyndsey Franklin, and Alex Endert. 2017. Warning, Bias May Occur: A Proposed Approach to Detecting Cognitive Bias in Interactive Visual Analytics. *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2017), 104–115. https://doi.org/10.1109/VAST.2017.8585669

[198] Emily Wall, Arpit Narechania, Adam Coscia, Jamal Paden, and Alex Endert. 2022. Left, Right, and Gender: Exploring Interaction Traces to Mitigate Human Biases. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2022), 966–975. https://doi.org/10.1109/TVCG.2021.3114862

[199] Emily Wall, John Stasko, and Alex Endert. 2019. Toward a Design Space for Mitigating Cognitive Bias in Vis. *2019 IEEE Visualization Conference (VIS)* (2019), 111–115. https://doi.org/10.1109/VISUAL.2019.8933611

[200] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), 1–15. https://doi.org/10.1145/3290605.3300831 Publisher: Association for Computing Machinery.

[201] P.C. Wason and J.ST.B.T. Evans. 1974. Dual processes in reasoning? *Cognition* 3, 2 (1974), 141–154. https://doi.org/10.1016/0010-0277(74)90017-1

[202] Martin Wenglinsky. 2017. A Meta-Scientific Perspective on "Thinking: Fast and Slow. https://www.wenglinskyreview.com/wenglinsky-review-a-journal-of-culture-politics/2017/1/23/kahnemans-fallacies. Accessed: 2024-11-08.

[203] Gregory Wheeler. 2020. Bounded Rationality. In *The Stanford Encyclopedia of Philosophy* (Fall 2020 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[204] Elizabeth Whitaker, Ethan Trewhitt, Matthew Holtsinger, Christopher Hale, Elizabeth Veinott, Chris Argenta, and Richard Catrambone. 2013. The effectiveness of intelligent tutoring on training in a video game. *2013 IEEE International Games Innovation Conference (IGIC)* (2013), 267–274. https://doi.org/10.1109/IGIC.2013.6659157

[205] Daniel T. Willingham. 2008. Critical Thinking: Why Is It So Hard to Teach? *Arts Education Policy Review* 109, 4 (2008), 21–32. https://doi.org/10.3200/AEPR.109.4.21-32 arXiv:https://doi.org/10.3200/AEPR.109.4.21-32

[206] Timothy D Wilson and Nancy Brekke. 1994. Mental Contamination and Mental Correction: Unwanted Influences on Judgments and Evaluations. *Psychological bulletin* 116, 1 (1994), 117.

[207] Timothy D. Wilson, David B. Centerbar, and Nancy Brekke. 2002. *Mental Contamination and the Debiasing Problem.* Cambridge University Press, 185–200.

[208] Luyan Xu, Mengdie Zhuang, and Ujwal Gadiraju. 2021. How Do User Opinions Influence Their Interaction With Web Search Results? *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (2021), 240–244. https://doi.org/10.1145/3450613.3456824 Publisher: Association for Computing Machinery.

[209] Yusuke Yamamoto and Yamamoto Takehiro. 2018. Query Priming for Promoting Critical Thinking in Web Search. *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (2018), 12–21. https://doi.org/10.1145/3176349.3176377 Publisher: Association for Computing Machinery.

[210] Mingzhe Yang, Hiromi Arai, Naomi Yamashita, and Yukino Baba. 2024. Fair Machine Guidance to Enhance Fair Decision Making in Biased People. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 285, 18 pages. https://doi.org/10.1145/3613904.3642627

[211] Liudmila Zavolokina, Kilian Sprenkamp, Zoya Katashinskaya, Daniel Gordon Jones, and Gerhard Schwabe. 2024. Think Fast, Think Slow, Think Critical: Designing an Automated Propaganda Detection Tool. *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024). https://doi.org/10.1145/3613904.3642805 Publisher: Association for Computing Machinery.

[212] Weiyu Zhang, Tian Yang, and Simon Tangi Perrault. 2021. Nudge for Reflection: More Than Just a Channel to Political Knowledge. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021). https://doi.org/10.1145/3411764.3445274 Publisher: Association for Computing Machinery.

[213] Yunfeng Zhang, Rachel K.E. Bellamy, and Wendy A. Kellogg. 2015. Designing Information for Remediating Cognitive Biases in Decision-Making. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015), 2211–2220. https://doi.org/10.1145/2702123.2702239 Publisher: Association for Computing Machinery.

[214] Yu Zhang, Jingwei Sun, Li Feng, Cen Yao, Mingming Fan, Liuxin Zhang, Qianying Wang, Xin Geng, and Yong Rui. 2024. See Widely, Think Wisely: Toward Designing a Generative Multi-agent System to Burst Filter Bubbles. *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024). https://doi.org/10.1145/3613904.3642545 Publisher: Association for Computing Machinery.

[215] Jason Chen Zhao, Wai-Tat Fu, Hanzhe Zhang, Shengdong Zhao, and Henry Duh. 2015. To Risk or Not to Risk? Improving Financial Risk Taking of Older Adults by Online Social Information. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (2015), 95–104. https://doi.org/10.1145/2675133.2685033 Publisher: Association for Computing Machinery.

[216] Chengbo Zheng, Yuheng Wu, Chuhan Shi, Shuai Ma, Jiehui Luo, and Xiaojuan Ma. 2023. Competent but Rigid: Identifying the Gap in Empowering AI to Participate Equally in Group Decision-Making. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023). https://doi.org/10.1145/3544548.3581131 Publisher: Association for Computing Machinery.

[217] Qian Zhu, Leo Yu-Ho Lo, Meng Xia, Zixin Chen, and Xiaojuan Ma. 2022. Bias-Aware Design for Informed Decisions: Raising Awareness of Self-Selection Bias in User Ratings and Reviews. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 496 (nov 2022), 31 pages. https://doi.org/10.1145/3555597

[218] Verena Zimmermann and Karen Renaud. 2021. The Nudge Puzzle: Matching Nudge Interventions to Cybersecurity Decisions. *ACM Trans. Comput.-Hum. Interact.* 28, 1 (2021). https://doi.org/10.1145/3429888

3.3 Article I Appendix

In addition to our scoping review, we identified and analysed the list of cognitive biases studied in our paper corpus. We include the analysis of the term usage and definition references in Article I. In this section, we categorise these biases according to their problem attribution to our real-world decision-making.

3.3.1 Problem Attribution of Cognitive Biases in HCI Studies

We categorised cognitive biases into four broad categories based on Benson's problem attribution of cognitive biases [11, 48]: *Too Much Information*, *The Need to Act Fast*, *Not Enough Meaning*, and *What to Remember*. These categories represent different ways humans overcome and simplify the challenges of everyday information processing – information overload, time scarcity, ambiguity, and limited memory – by applying heuristics and cognitive biases. Table 3.1 provides a comprehensive list of cognitive biases categorised by their problem attribution. In the following, we discuss each category of cognitive biases from the angles of HCI and interaction design.

- **Too Much Information.** 18 cognitive biases in our corpus (covering 75 articles) fall into this category. Information overload is a familiar issue in HCI [8, 134], as user interfaces tend to overwhelm users with information. (e.g., only 32% of the total available features on Microsoft Word are commonly used[1]). To overcome this challenge, humans apply different cognitive biases to filter out information, such as confirmation bias, anchoring bias, or the framing effect.

- **The Need to Act Fast.** 23 cognitive biases (covering 29 articles) are attributed to this problem. Humans are subject to making decisions under time pressure. Several kinds of user interfaces pressure users to expedite their decisions [116] (e.g., "only 2 more seats are available" on flight booking platforms). To make fast (and frugal) decisions, users resort to cognitive biases like loss aversion bias or the ambiguity effect (tendency to avoid options with uncertainty).

- **Not Enough Meaning.** 15 cognitive biases (covering 17 articles) are explained by this problem attribution. To make sense of the world, people tend to connect the dots and generalise from sparse, incomplete data. In HCI, users seldom have complete information about the systems' capabilities because they are black boxes [41]. Therefore, they rely on certain cognitive biases like the placebo effect (the tendency to judge something as efficient because of one's prior expectations) or the bandwagon effect (the tendency to do something primarily because others are doing it).

- **What to Remember.** 12 cognitive biases (covering 15 articles) are related to this problem. Humans are selective in retaining information because they have limited memory. User interfaces like search results pages place ads or profitable items at the top of the results to make them stick out in users' memory. Consequently, users express cognitive biases like the priming effect or the peak-end rule (tendency to more easily remember emotionally intense moments, including the end of an event).

- **Other Biases.** 18 cognitive biases (covering 15 articles) do not fall into any problem attribution, for example, cognitive dissonance, and uncertainty bias. This set of biases does not include non-specific cognitive bias, which appears in 20 articles.

[1] https://answers.microsoft.com/en-us/msoffice/forum/all/what-is-the-percentage-of-features-of-word-excel/80e417ef-8336-49a5-9f5f-0a59c8c8fbd4

Table 3.1: A comprehensive list of cognitive biases identified in our scoping review (Article I) with their problem attribution based on Benson's cognitive bias codex [11]. Biases with synonyms were merged together; for example, the term *Anchoring Effect* was merged with the more frequently used *Anchoring Bias*. (Refer to Table 1 in Article I for synonyms)

Problem Attribution	**Cognitive Biases Investigated in HCI Studies**
Too Much Information (19 biases in 75 articles)	Confirmation Bias, Framing Effect, Anchoring Bias, Availability Bias, Default Bias, Omission Bias, Motivated Reasoning, Weber-Fechner Law, Contrast Bias, Mere-Exposure Effect, Disconfirmation Bias, Continued Influence Effect, Vividness Criterion, Contrast Effect, Conservatism, Blind Spot Bias, Choice Supportive Bias, Salience Bias
Need To Act Fast (23 biases in 29 articles)	Decoy Effect, Loss Aversion Bias, Illusion of Control, Status-Quo Bias, Risk Aversion Bias, Sunk Cost Fallacy, Overconfidence Bias, Hyperbolic Discounting Effect, Ambiguity Aversion, Dunning-Kruger Effect, Illusion of Explanatory Depth, Backfire Effect, Premature Closure, Less-Is-More Effect, Optimism Bias, Forer-Barnum Effect, Self-Other Bias, Attraction Effect, IKEA Effect, Fundamental Attribution Error, Belief Bias, Commitment Bias, Regret Aversion Bias
Not Enough Meaning (15 biases in 17 articles)	Bandwagon Effect, Automation Bias, Representativeness Heuristics, Placebo Effect, Halo Effect, Reinforcement Effect, Exposure Effect, Planning Fallacy, Selective Perception, Gambler Fallacy, In-Group Bias, Spotlight Effect, Narrative Fallacy, Illusory Correlation, Out-Group Bias
What To Remember (12 biases in 15 articles)	Priming Effect, Order Effect, Peak-End Rule, Implicit Associations, Attention Bias, Primacy Effect, Recency Bias, Position Bias, Negativity Bias, Fading Affect Bias, Positioning Heuristics
Other Biases (18 biases in 15 articles)	Cognitive Dissonance, Scarcity Bias, Ranking Bias, Illusion Bias, Scarcity Effect, Control Bias, Uncertainty Bias, Decision Fatigue, Coping with Evidence of Uncertain Accuracy, Oversensitivity to Consistency, Absence of Evidence, Present-Biased Preferences, Bayesian Reasoning, Worker Bias, Affect Heuristic, Self-Fulfilling Prophecy, Reciprocity Bias, Domain Bias, Approach Bias

3.3.2 How Do Computing Systems Trigger Cognitive Biases in Users?

Our findings suggest that cognitive biases can be triggered in a variety of circumstances. This phenomenon is in parallel with the existing surveys of cognitive biases in explainable AI [14, 101, 189]. By categorising cognitive biases in our corpus using the Benson Cognitive Bias Codex [11], which attributes cognitive biases to fundamental challenges that humans face, we show that the majority of studied cognitive biases in our corpus can be attributed to the problems of *Too Much Information* and *The Need to Act Fast*. In other words, users of computing systems employ heuristics to simplify complex information and expedite decision-making. From the system perspective, we can imply that **computing systems tend to *overwhelm* users with information and *restrict* time to make decisions**. This is in line with the current landscape of HCI, which tends to promote seamless and efficient interactions [59, 77, 79] while presenting users with a large amount of mentally demanding information [8, 134]. Prominent scenarios of such interactions include social media platforms, which tend to overburden users with information [107, 164] and offer them features to rapidly skim through information (e.g., short-form videos [27]). Moreover, our review suggests other cognitive biases spanning from the two other problem attribution: *What to Remember* and *Not Enough Meaning*. This also

implies that computing systems have the potential to constrain users' *memory capacity* and lead to *ambiguity* in the interaction. Nonetheless, we envision future research to empirically investigate the suggested phenomena in HCI in relation to cognitive biases: how do information overload, time constraint, short memory, and ambiguity present in the interaction with computers affect the manifestation of cognitive biases?

3.4 Chapter Reflection

Humans employ cognitive biases and heuristics as mental strategies to make sense of the world and efficiently form decisions. Specifically, we face constraints for making everyday decisions because we neither have sufficient cognitive bandwidth, memory capacity, time to act, nor complete knowledge of the world. Despite the ongoing debates about the definitions of cognitive biases, the scientific community widely accept that cognitive biases (as coined by Tversky and Kahneman [178]) happen naturally without one's awareness and systematically influence human judgment and behaviour. However, we lack sufficient understanding of the occurrences of these biases during human-computer interaction. Specifically, the interdisciplinary nature of HCI suggests that researchers study cognitive biases using different methodologies and with different focuses and application contexts. Yet, we do not have a clear grasp of how biases are studied in HCI.

We address **(RQ 1)** through a scoping review of 127 HCI articles that study cognitive biases. We identify different ways HCI researchers engage with cognitive biases: effect study, mitigation, utilisation, quantification, and observation. The findings then outline the three-layer narratives of cognitive biases in HCI: computing systems can trigger and mitigate the effects of cognitive biases; designers can capitalise on users' cognitive biases and steer their behaviours; and HCI researchers develop tools and methods to better observe these phenomena. We leverage these narratives into a summary of how HCI researchers study cognitive biases (Figure 1.1) – computing systems trigger users' existing cognitive biases and influence their behaviours in the real world. This motivates HCI researchers to build tools and methods to understand the manifestation of cognitive biases in HCI and derive designs of computing systems and user interfaces that consider these biases, mitigate their undesired effects, and leverage useful biases.

Based on our analysis, we discuss the role of cognitive biases in HCI as a double-edged sword. While computing systems can trigger and amplify cognitive biases, which lead to devastating outcomes, systems can leverage cognitive biases in a meaningful way through digital nudging, which taps into users' cognitive biases to change their behaviours (e.g., to encourage healthy eating habits). However, we acknowledge the ethical considerations arising from using cognitive biases in people. Designs that capitalise on cognitive biases bear the consequences of manipulating people's decision-making and harming their autonomy. We call for designers to account for potential harms arising from cognitive biases that could emerge during the interaction and give the users transparency and awareness about their biases being used.

Our scoping review reflects on the research conduct around the issue of cognitive biases in HCI. Particularly, HCI articles use a variety of terminology and definitions that refer to cognitive biases. At the same time, precise terminology and definitions are sometimes absent in the literature. The field of HCI largely borrows definitions from neighbouring fields; therefore, we signal the need for the HCI community to better connect with research discourses in the originating fields of behavioural science and psychology. While our scoping review provides the view of human-computer interaction through the lenses of Tversky and Kahneman's cognitive biases, the scope of this review can restrict the understanding of systematic effects in HCI beyond cognitive biases. There are psychological constructs that can be considered cognitive biases, such as design fixation, selective exposure, or decision-making fairness. We call for future investigations into the broader systematic effects arising from the human mind in HCI.

To conclude this chapter, we pinpoint the issue of cognitive bias as an emerging area in HCI research. There are potential improvements in understanding and opportunities for cognitive bias research in HCI.

Specifically, there are limited investigations into tools and methods to quantify the effects of cognitive biases in HCI. In the next chapter, we discuss the design of experiments to elicit cognitive biases and the potential of physiological sensing as a tool to detect the effects of cognitive biases in situ.

Chapter 4

Quantifying the Occurrences of Cognitive Biases

4.1 Introduction

Cognitive biases influence how individuals perceive and process information. Informed by psychology, people rely on heuristics to help sift through the complexity of the information. These heuristics are highly dependent on individuals' preferences and existing beliefs. One form of heuristics is when people rely on the *congruence* between their beliefs and the information content's stance. This heuristic is especially prominent in the context of interacting with online information and social media, which tends to contain polarising, opinionated messages. In particular, the reliance on one's beliefs gives rise to the application of mental shortcuts and cognitive biases. The literature, as well as real-world examples, show that these biases exacerbate the creation of echo chambers and the spread of misinformation. Facing polarising information on the Internet, individuals tend to rely on their *subjective* beliefs rather than objectively assess the stance of the content. However, such biased tendencies generally happen without the user's awareness.

In the previous chapter, we highlight limited research on quantifying the effects of cognitive biases, which is a crucial step towards effectively mitigating biases. The literature suggests that one should make sure that cognitive biases actually occur before applying interventions. With the ability to detect the occurrences of cognitive biases in the interaction, designers can capture the user states and the interaction contexts where biases manifest and prompt interventions to address specific scenarios. However, one major challenge is obtaining a reliable ground truth for the occurrences of cognitive biases. Prior research investigates methods to quantify the occurrences of cognitive biases, including self-reports and behavioural measures. However, these measurements do not accurately indicate the effects of cognitive biases. Specifically, it is infeasible to measure cognitive biases from self-report measures because individuals are unaware of their own biases. Moreover, self-presentation and falsification issues could confound the responses. Behavioural measures, such as dwelling time or eye-tracking data, are prone to ambiguity. Research suggests that such measures produce mixed results.

Physiological measurements offer more objective and non-intrusive probes into human cognitive states and, therefore, present promising means to quantify the occurrences of cognitive biases. Specifically, research in neuropsychology employs physiological signals to study human information processing, as they reflect how our brains and bodies respond to and process information stimuli. Prior research employs electrodermal activity (EDA) and functional magnetic resonance imaging (fMRI) to study a psychological construct called *dissonance arousal*, which manifests itself in the form of physiological discomfort. It is also associated with cognitive dissonance, which occurs when individuals simultaneously hold conflicting beliefs.

In this chapter, we present two empirical user studies to explore indicators for the occurrences of cognitive biases when individuals read different opinions. In both studies, we exposed participants to a series of text and image stimuli containing opinions on a polarising topic (e.g., climate change, abortion rights, or feminism). These opinions express statements that support one side of the ideological spectrum. For example, on the abortion rights topic, an opinion can either support the idea that "abortion should be legal" or the idea of "abortion should be illegal." Therefore, we assume that these opinions are either *congruent* or *dissenting* with the participant's ideological beliefs, which we gauged prior to the study. During exposure to ideological polarising text stimuli, we apply three classes of measurements: self-report, behavioural, and physiological. Specifically, we ask participants three self-report questions that reflect their perception of each stimulus. As behavioural measures, we measure fixation, saccade, and dwelling time through an eye-tracking camera. As physiological measures, we apply physiological sensors to measure hemodynamic activity (brain oxygenation levels) and electrodermal activity (EDA) using wearable functional near-infrared spectroscopy (fNIRS) and wristband EDA electrodes.

The designs of both studies are similar. The first study exposed participants to both image and text stimuli and presented each stimulus continuously in a randomised order. The findings are, however, inconclusive because the study design did not provide a time gap between the two stimuli. Therefore, physiological responses may not reflect the reactions induced by the stimulus but, instead, by the preceding stimuli. Physiological measurements, especially hemodynamic activity, have a few-second delay for event-related responses. Therefore, we revised the design of the first study into the second study by minimising the confounds of the first study. Specifically, we provided an inter-stimulus interval (ISI) between two consecutive stimuli to observe clearer changes in the physiological measurements. We also removed image stimuli as they are highly contextual (i.e., depending on how they are presented with text information) and, thus, can be ambiguous.

The findings suggest that physiological expressions can be reliable indicators of the occurrence of cognitive biases when individuals read different opinions. In the second study, we found significant effects of ideological congruence on hemodynamic activity. However, such effects were only pronounced in individuals with low interest in the topic. We also observed non-significant trends for the effects of ideological congruency on the skin conductance responses (derived from EDA). Therefore, the implication of this study is two-fold: it shows that not only are physiological measurements a means to quantify the effects of cognitive biases, but also user-related factors can influence these biases. The findings of both studies provide empirical and methodological contributions to HCI. We demonstrated a study design that induces cognitive biases, employed physiological measures to monitor their effects in situ, and cross-validated the findings with self-report and behavioural measures. Further, we present a first step to building *bias-aware systems* as computing systems that detect and consider the presence of cognitive biases in users. We describe in more detail the two studies and their implications in Article II.

4.2 Article II

This article was presented at the CHI Conference on Human Factors in Computing (CHI 2023). It received the honourable mention recognition for the best paper (top 5% of the total submissions).

Nattapat Boonprakong, Xiuge Chen, Catherine E. Davey, Benjamin Tag, and Tilman Dingler. 2023. Bias-Aware Systems: Exploring Indicators for the Occurrences of Cognitive Biases when Facing Different Opinions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3544548.3580917

Ethics Application ID: 1956072.1, the University of Melbourne Human Research Ethics Committee.

Bias-Aware Systems: Exploring Indicators for the Occurrences of Cognitive Biases when Facing Different Opinions

Nattapat Boonprakong
nboonprakong@student.unimelb.edu.au
The University of Melbourne
Melbourne, Victoria, Australia

Xiuge Chen
xiugec@student.unimelb.edu.au
The University of Melbourne
Melbourne, Victoria, Australia

Catherine E. Davey
catherine.davey@unimelb.edu.au
The University of Melbourne
Melbourne, Victoria, Australia

Benjamin Tag
benjamin.tag@unimelb.edu.au
The University of Melbourne
Melbourne, Victoria, Australia

Tilman Dingler
tilman.dingler@unimelb.edu.au
The University of Melbourne
Melbourne, Victoria, Australia

ABSTRACT

Cognitive biases have been shown to play a critical role in creating echo chambers and spreading misinformation. They undermine our ability to evaluate information and can influence our behaviour without our awareness. To allow the study of occurrences and effects of biases on information consumption behaviour, we explore indicators for cognitive biases in physiological and interaction data. Therefore, we conducted two experiments investigating how people experience statements that are congruent or divergent from their own ideological stance. We collected interaction data, eye tracking data, hemodynamic responses, and electrodermal activity while participants were exposed to ideologically tainted statements. Our results indicate that people spend more time processing statements that are incongruent with their own opinion. We detected differences in blood oxygenation levels between congruent and divergent opinions, a first step towards building systems to detect and quantify cognitive biases.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; *Empirical studies in HCI*; *Ubiquitous and mobile computing systems and tools.*

KEYWORDS

Bias-aware systems, Cognitive biases, Cognition-aware systems, fNIRS, Eye tracking, Electrodermal activity

ACM Reference Format:
Nattapat Boonprakong, Xiuge Chen, Catherine E. Davey, Benjamin Tag, and Tilman Dingler. 2023. Bias-Aware Systems: Exploring Indicators for the Occurrences of Cognitive Biases when Facing Different Opinions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3544548.3580917

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3580917

1 INTRODUCTION

Algorithms increasingly curate the information we encounter online. In an attempt to grab and keep users' attention, they filter and provide content based on prior browsing history and inferred interests [9, 46]. Consequently, most information provided to users feeds into their existing beliefs and opinions. In recent years, this mechanism has triggered a heated discussion about how the prioritisation of user engagement plays into the spread of misinformation and political extremism [41]. While algorithms have been shown to be attributing factors, users themselves seem to process information differently based on their pre-existing notions and beliefs [42, 61].

Facing vast amounts of information online, people adopt cognitive strategies to filter and sift through content more effectively. Such behaviour fosters the occurrence and application of what is referred to as *cognitive biases*, i.e., mental shortcuts we take while processing information. Personal preferences and prior experiences play heavily into this simplification of information processing by focusing on the known or familiar [102].

Misinformation tends to thrive in an environment of simplification and repetition. Its spread, prevalence, and persistence have had real-world implications, such as negative health impacts. For example, the belief in a link between vaccinations and autism has led to parents withholding crucial immunisation from their children resulting in the return of preventable diseases [83]. Misinformation about the dangers and risks of vaccinations keep influencing public debates about the effectiveness of COVID-19 measures to this date [58].

Misinformation is further fueled by frequent exposure. What we encounter more often appears more familiar and can be falsely attributed to a certain truism. When Weaver and colleagues [108] repeatedly showed study participants the same statement from the same communicator, for example, participants perceived the general consensus on that statement to be greater the more often they encountered it. Hence, systems, websites, and platforms that cater to our interests and beliefs tend to skew our perceptions and amplify our innate cognitive biases [7]. This becomes an even bigger problem when it affects our decision-making and opinion formation in the real world, such as on topics like climate change, immigration policies, or abortion rights.

Especially such polarising topics often lead to the segregation of like-minded people. Echo chambers and filter bubbles are two well-known phenomena that contribute to one-sided information

exposure and the spread of misinformation. They capitalise on people's biases, most and foremost on what is referred to as confirmation biases [3, 45, 62, 81]. This bias is expressed in people's tendencies to seek out and favour information that aligns with their existing beliefs and expectations while ignoring dissenting information [56, 75]. While it is crucial to mitigate the negative effects of cognitive biases, we first have to understand when and in what situation biases occur, what triggers them, and how they can be reliably *quantified*.

Researchers have examined behavioural measures for exposing confirmation bias or what Klapper called *selective exposure*, i.e., the tendency to seek out predominantly information that supports one's beliefs [51]. This effect has been demonstrated to be present in news dwelling time [37], web browsing behaviour [54, 103], and eye-tracking information [68, 91, 99]. Behavioural measures provide an unintrusive way of tracking selective exposure [20]. Yet, these measures have produced mixed results and interpretations. For instance, researchers used dwelling time as an indicator of confirmation bias as studies have shown that users spend more time reading congruent information and less time on dissenting information [37, 68]. Meanwhile, some research found a rather opposite effect as users spent more time reading attitude-challenging opinions [37, 100]. At the same time, Sülflow et al. [99] and Zillich and Guenther [112] reported no significant differences in reading time between congruent and dissenting information.

A major difficulty in researching cognitive biases is obtaining reliable ground truth for their occurrence. While we could simply ask users whether they have exhibited biased information consumption behaviour, self-report responses are not always reliable since they may be confounded by a broad range of factors, like self-presentation [96] and preference falsification [59]. Recent research has investigated the use of physiological sensors in evaluating cognitive biases [74, 76, 105, 110]. Physiological signals have been regarded as the (more) objective means to quantify mental states [38]. They reflect how our brains and bodies respond and process information [104]. Although physiological signals may be objective measures of our innate cognitive biases, it is unclear how biases manifest themselves in physiological data or can be effectively measured.

In this work, we focus on whether *physiological signals can be a reliable, objective measure of cognitive biases* in an attempt to equip computing systems with the ability to detect and eventually help users mitigate them. We were specifically interested in the occurrence of cognitive biases while processing information that is either congruent with or diverges from people's existing beliefs. Hence, we conducted two studies in which we exposed participants to stimuli that represented an ideologically congruent opinion and those depicting a dissenting opinion. Throughout these studies, we recorded behavioural and physiological signals, such as eye movement data, electrodermal activity, and brain oxygenation levels (via fNIRS) along with self-reports to explore physiological and behavioural expressions indicating the congruence between users' opinions and the presented statements. We also investigated the interplay between the manifestations of biases and the individuals' interest and familiarity with the topic.

Our results show that participants tended to spend more time but less reading effort on ideologically dissenting stimuli. We also found that topic interest significantly impacted the effects of opinion congruency: especially individuals with low interest in a topic exhibited higher neural activity when they were exposed to attitude-dissenting information. Through this work, we contribute the following:

- We present two studies aiming to explore how cognitive biases manifest themselves in behavioural and physiological signals by presenting ideologically polarised statements and recording physiological and interaction data as well as self-reports.
- Based on our findings, we discuss the notion of *bias-aware systems* – i.e., computing systems that detect and take into account the presence of cognitive biases in users – and their potential to detect, quantify, and mitigate the effects of cognitive biases. We discuss challenges, opportunities, and ethical considerations for bias-aware systems from what we learned from this research.

2 BACKGROUND

Our work is mainly grounded in research on behavioural psychology and psychophysiology while touching on recent discussions in human-computer interaction [25–27] regarding the unintended effects of cognitive biases in users.

2.1 Cognitive Biases

Cognitive biases refer to a systematic pattern of deviation from norm or rationality in judgement [39]. The concept was proposed in the work of Tversky and Kahneman in 1974 [102]. Tversky and Kahneman explained different types of heuristics, or so-called *mental shortcuts*, employed by humans to avoid overwhelming their limited cognitive resources by preferably using automatic thinking (System 1) over rational thinking (System 2) [48]. While heuristics enable us to reach a decision faster, they become problematic as they generally distort our rationality in ways we are unaware of.

When making decisions or judgments, individuals who exhibit cognitive biases tend to follow their own beliefs or preferences rather than objective information [39]. In the context of information consumption, this leads to a distortion of the way people perceive and evaluate information, often resulting in favouring information that supports their attitudes [47]. Cognitive biases can be present in many forms. Prominent examples include confirmation bias (seeking predominantly information that aligns with one's beliefs [75]), cognitive dissonance (avoiding information that conflicts with one's beliefs [30]), or negativity bias (responding to negative stimuli with stronger attention and emotional responses [52]). Confirmation bias and cognitive dissonance, for example, are potential contributors to selective exposure [72, 95, 96]. This describes the tendency to seek out predominantly information that supports one's beliefs or attitudes while avoiding dissenting information [51]. This impacts how critical people evaluate information [80, 113] and potentially fosters ideological polarisation [54, 97]. A prominent example in the 20th century was the use of one-sided news reporting by the German government in the 1930s and 40s. Consequently, the belief systems of the majority of Germans who grew up under the regime were skewed towards anti-semitism [106]. In a similar, but less extreme fashion, the recent examples of vaccine hesitency [28]

and climate change denial [70] have shown that confirmation bias limits informed and objective discussions of polarizing topics.

People tend to save up their limited cognitive resources when processing information, which makes them vulnerable to various types of manipulation [39]. Today, new information is continuously available to people, which results in excessive mental demand, or mental overload. To prevent overexerting their cognitive resources, people employ cognitive biases or "mental shortcuts" to simplify the complexity and filter out the most relevant information. This is expressed in making faster but less deliberate decisions [48].

Together, cognitive biases and personalised recommendation algorithms contribute to the formation of filter bubbles through a reinforcing loop [3, 62]. When exploring information online, users exhibit their cognitive biases by selectively exposing themselves to certain types of information. Meanwhile, recommendation algorithms detect patterns in the selective consumption of information and optimise themselves to keep engagement high by catering predominantly to what the users prefer [9, 46]. Consequently, the users' innate biases are further amplified. In sum, recommendation systems and selective exposure build a self-reinforcing loop: the former curate content items that are congruent with the users' preferences; at the same time, users seek and favour such content due to confirmation bias [3]. In other words, cognitive biases in individuals can be reinforced by automated recommendation systems.

2.2 Two-step Model of Processing Conflicting Information

While cognitive biases often manifest when facing different opinions, their occurrences also depend on the prior background of the information consumer. In his series of works, Richter [85–87] proposes a two-step model of validation. The model states that people tend to use the perceived plausibility of the information as their heuristics. When encountering information, people first employ *Epistemic Monitoring* to evaluate whether the content is compatible with their beliefs or preferences. In general, people save up their cognitive resources by allocating them to information that is congruent with their beliefs. This results in people processing information with cognitive biases. However, individuals with higher working memory resources, advanced epistemological beliefs, or relevant background knowledge may pursue the second step – *Elaborative Processing* – at which they process the information in a more balanced and objective manner.

2.3 Quantifying the Effects of Cognitive Biases

Being able to quantify the occurrence and the effects of cognitive biases comes with numerous benefits [65]. With the awareness of the users' biases, interventions can be designed to help users overcome their irrationality and become more critical and deliberate when facing information online. However, given that cognitive biases normally happen without people being aware of them, it is challenging to objectively define and measure them [5]. In this section, we review methodological approaches to quantifying the effects of cognitive biases in the context of information consumption, using behavioural measures and physiological signals.

2.3.1 Behavioural Measures. Recent research in the field of selective exposure has used behavioural measures, i.e., through direct observations or in-lab studies [20]. By exposing users to attitudinal information, researchers were able to observe the deviation of users' behaviour as by-products from the manifestation of their innate cognitive biases. Commonly, researchers have used measures like dwelling time – i.e., the amount of time participants exposed themselves to certain types of information – and information selectivity (e.g., the number of content clicks or page visits). Recent approaches have utilised eye tracking measures as they offer advantages over dwelling time, for example, more insights into the users' visual attention [68, 99].

While behavioural expressions offer an unobtrusive measure of bias, research that employed behavioural measures has produced mixed results. Some works showed that people tended to spend more time on what confirms their opinions [68, 91]. Marquart [68], for example, tracked fixation time in online news reading and found that people tended to spend more time with news items that were compatible with their beliefs. Meanwhile, some studies suggested a rather opposite phenomenon [37, 100]. Taber and Lodge [100] found that individuals spent significantly longer time reading attitude-challenging arguments. Some works reported no significant deviation in dwelling time [99, 112]. For instance, an eye-tracking study by Sülflow et al. [99] suggested no effects of opinion congruency on the users' attention to social media news posts but found higher selectivity for attitude-reinforcing contents.

2.3.2 Physiological Measures. Given that our innate biases are the consequence of the interplay of the complex regulation of our cognitive and affective states, cognitive biases are likely to induce physiological changes. Research has long investigated the effects of cognitive dissonance on human physiology. Since the introduction of cognitive dissonance by Festinger [30], a series of studies have investigated a psychological construct called *dissonance arousal* which manifests itself in the form of physiological discomfort [111].

Research by Westen et al. [110] has probed the presence of cognitive biases using physiological signals. Westen and colleagues used functional Magnetic Resonance Imaging (fMRI) to assess the effects of cognitive dissonance and found significantly higher neural activations when the users were processing ideological dissenting information. Subsequent works have confirmed such findings [12, 49]. Meanwhile, Ploger et al. [82] used electrodermal activity (EDA) and heart rate to assess dissonance arousal by exposing individuals to video clips that present attitude-challenging information. However, they found weak effects from ideological (in)congruency.

2.4 Physiological Signals

Physiological signals have been widely used as a surrogate to measure cognitive states [14]. They reflect the reactions from our brains and bodies through a variety of signals. In our work, we focus on two particular signals: electrodermal activity, a widely used physiological measure in HCI, and hemodynamic responses, a non-invasive way to measure brain activities.

2.4.1 Electrodermal Activity. EDA refers to the variation of the electrical conductance of the skin [8], which results from the skin's sweating function. The changes in the sympathetic nervous system control the level of sweating on the skin and thus the EDA. The signal is often collected from electrodes placed on specific body

parts, for example, on the fingers or the wrist. In the HCI community, EDA is known as a low-cost, unobtrusive physiological measure [4, 22].

EDA consists of two signal components: Skin Conductance Responses (SCR) and Skin Conductance Level (SCL). SCR represents high-frequency, short-term spikes in the EDA signal triggered by eliciting stimuli. SCL denotes inertial, long-term changes in the EDA. Researchers have used EDA as a marker for negative cognitive activity, for example, cognitive workload [57, 93] and arousal [22, 35, 67].

2.4.2 Hemodynamic Responses. To quantify hemodynamic responses or the changes in blood flow to the brain, researchers have used functional Magnetic Resonance Imaging (fMRI) and functional Near-Infrared Spectroscopy (fNIRS) to infer the relative changes in the concentration of oxygenated haemoglobin ([HbO]) and deoxygenated haemoglobin ([HbR]) [16, 50]. Since haemoglobin absorbs near-infrared light, one can derive the haemoglobin concentration as a function of optical density [6]. Greater changes in haemoglobin concentration are associated with higher levels of neural activation. Therefore, fMRI and fNIRS offer a measurement of innate neural activity [92].

Unlike fMRI, fNIRS provides a less invasive and more noise-robust method to monitor the hemodynamic responses and, thus, the brain activity [66]. Recent research has employed fNIRS to assess a variety of psychological constructs, for example, cognitive workload [2, 31] and affective states [40, 43].

2.5 Summary

Our biases are especially problematic when they come into play for nuanced discussions on polarised topics. They are often exacerbated by the way we consume information online. While they serve us when sifting through vast amounts of information, they at times compromise our ability to make objective decisions. Prior research has investigated how to "track down" the presence of cognitive biases by studying their effects on behavioural measures. While dwelling time as a behavioural measure may be an indicator of cognitive biases, it has been shown to not always be reliable. Recent research has investigated the use of physiological responses to probe the effects of bias. In the context of information consumption, researchers have used fMRI, EDA, and heart rate to observe such effects. Our work adds up to the literature by using physiological signals to monitor the presence of cognitive biases when exposed to opinions from different ideological spectra. To the best of our knowledge, our work is the first to apply fNIRS signals to study the effects and occurrences of biases in the context of information exposure with the intent to study the notion of bias-aware computing systems. In the following, we present two studies, in which we expose participants to a range of opinions and probe their interactions, behavioural expressions, and physiological data to explore how cognitive biases may manifest themselves.

3 STUDY 1: DESIGN

We conducted Study 1 to explore different indicators for the occurrences of cognitive biases. In this study, we exposed participants to textual and image stimuli that represented opinions on four polarising topics. At the same time, we recorded behavioural data (eye tracking) and physiological signals, namely electrodermal activity (EDA), and brain hemodynamic responses using functional near-infrared spectroscopy (fNIRS).

3.1 Stimuli Selection

We operationalised stimuli that consisted of information on either end of the ideological spectrum, i.e., supporting information (pro) or contradicting information (con). Adapting to the Australian context, where this study was conducted, each stimulus was chosen with regard to ideologically polarising topics that were dominant in the current, domestic public debate. Consequently, we selected the following four topics for the study: *political progressivism*, *climate change*, *feminism*, and *multiculturalism in Australia*. All four topics were widely discussed in the media, and well-known to the Australian public with increasingly polarised viewpoints. Thus, we expected that the stimuli would have the potential to trigger strong attitudes and prompt cognitive biases in the study participants. Table 1 gives an overview of the pro-stances and con-stances for each of the four topics.

We selected **progressivism** due to the increasing ideological polarisation between progressive and conservative politics[1] since the 1970s [18, 19, 71, 109]. Similarly, we chose **climate change** because of the increasing discrepancy between those who acknowledge man-made climate change as opposed to denying it [64]. We also considered **multiculturalism** due to the lasting conflict between multiculturalism in Australia and the Anglo-Saxon inheritance rooted in the "White Australia" policy [21]. Lastly, **feminism** was selected because of the increasing pushback against feminism among Australian male groups [88] and third-wave feminists [98].

We used two types of stimuli: texts and images. Text stimuli were curated from either user opinions on Twitter[2] or the Procon.org website[3]. The latter source hosts information on both ends of the ideological spectrum, i.e., pros and cons, on different topics. We sourced climate change and feminism stimuli from Procon.org; progressivism and multiculturalism stimuli were curated from tweets posted in Australia from June to July 2021. We controlled all text stimuli for being in English and approximately 50 words in length.

While statements on ProCon.org are heavily contextualised to US politics and society, topics of global interest and the general discourse, such as climate change and feminism, are also applicable to the Australian context. We selected statements that do not contain US-specific information, e.g., excluding those mentioning US laws.

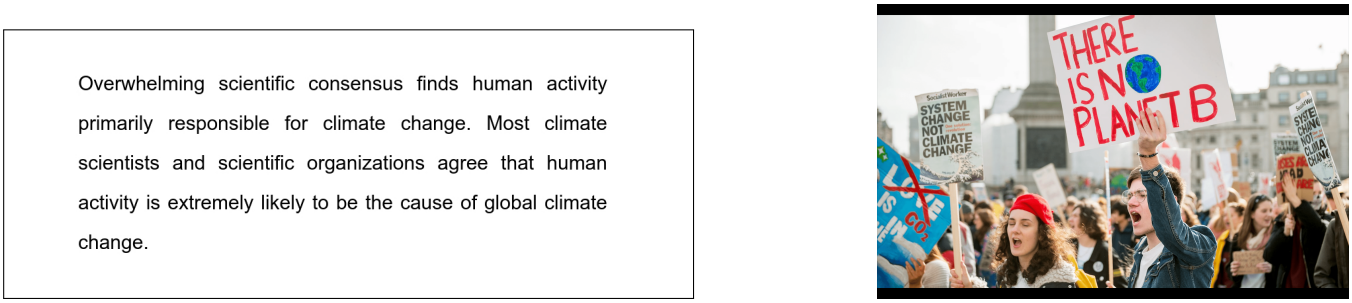
Each image stimulus was selected from online images or graphics that contained messages supporting an ideological viewpoint. Similarly, we picked those images from the ProCon.org website or keyword searches on Twitter. Examples of image stimuli were the cover of the book "The Greatest Hoax" [44] or a photo of a protest against man-made climate change.

We accumulated a total of 64 stimuli consisting of 32 texts, and 32 images. Stimuli presenting pros and cons were even in numbers. We presented participants with each stimulus on a screen. Every text stimulus was displayed in a single paragraph with the same font

[1] To avoid confusion among our readers, we use the Anglo-Saxon nomenclature. In Australia, conservative politics are actually called "liberal", whereas the Australian "labor" is the equivalent to the Anglo-Saxon progressives.
[2] www.twitter.com
[3] www.procon.org

Table 1: Topics and their ideological ends, as used in Study 1

Topic	Pro stance	Con stance
Political Progressivism	I support a political and societal change	I do not support political and societal change
Climate change	I believe humans are primarily responsible for climate change	I believe humans are not primarily responsible for climate change
Multiculturalism in Australia	I support multiculturalism in Australia	I support the Anglo-Saxon national identity of Australia
Feminism	I support feminism and women's rights	I do not support feminism and women's rights



(a) Text **(b) Image**

Figure 1: Examples of stimulus presentation for Study 1. Both stimuli were on the topic of climate change.

(Arial 30 px, black colour), line spacing (double), alignment (justified and centred), column width (800 px), and white background. Image stimuli were presented in an 800px × 800px resolution. Figure 1 shows an example of the stimuli used in Study 1.

3.2 Study Protocols

3.2.1 Experimental design. We studied the effects of the congruency of ideological stances between the user and the stimulus. To do so, we conducted two experiments with a 2-level (*Congruent* and *Dissenting*) within-subjects design: one to examine text stimuli and one to examine image stimuli. The *congruent* condition implied the stimulus' stance was aligned with the user's stance. Conversely, the *dissenting* condition implied the stimulus' stance contradicted the user's stance. Table 2 shows a list of independent and dependent variables of this study.

3.2.2 Participants. Through the university network, we invited 33 native or bilingual English speakers (19 women, 14 men) to participate in Study 1. The mean age of our participants was 32 (SD = 11.43) years. The minimum and maximum ages were 18 and 54, respectively. Of our participants, 15 possessed a postgraduate degree, 10 held a bachelor's degree, and the remaining six participants had at least year 12 education.

3.2.3 Procedure. The study took place in a quiet room. We informed each participant about the purpose and procedure of the study. After providing their consent in writing, we seated participants in a comfortable position and asked them to adjust their seats so that their heads were centred and approximately 60-65cm away from the monitor screen. We then asked each participant to respond to the pre-study survey and calibrated the placements of the physiological sensors.

We subsequently asked participants to read a series of text and image stimuli on a 24-inch monitor. After each stimulus, participants were asked to press the space key to proceed to the next one. The order of stimuli presentation was counterbalanced: participants either completed image stimuli first then text stimuli, or vice versa. Moreover, the order of the four topics was counterbalanced. Within each topic, stimuli were displayed in random order with no gap in between. Once a participant finished all stimuli for a topic, we paused the data collection for approximately one minute; then, participants continued reading the stimuli on the following topic. Upon completion, participants responded to a post-study survey and received a $20 voucher for compensation. The whole study took 45-60 minutes.

3.2.4 Sensors. Throughout the study, we recorded participants' eye movements, EDA, and hemodynamic responses. Eye movements were recorded with a Tobii Pro X3-120 eye tracker[4] with a sampling rate of 120 Hz. We mounted the eye tracker at the bottom of the monitor. We used the Empatica E4 wristband[5] to gather EDA data. To prevent potential motion artefacts, we asked participants to wear the wristband on their non-dominant hand. Additionally, we recorded functional near-infrared spectroscopy (fNIRS) from the participant's forehead using the BIOPAC fNIR Sensors 2000[6].

[4]https://www.tobiipro.com/product-listing/tobii-pro-x3-120/
[5]https://www.empatica.com/research/e4/
[6]https://www.biopac.com/product/fnir-sensors-2000/

Table 2: Summary of Independent and Dependent Variables in Study 1

Variables	Measures	Scale
Independent Variables	Participant-Stimulus Ideological Congruency	2 levels (congruent and dissenting)
Dependent Variables	Dwelling Time	Continuous
	Number of Fixations	Number of occurrences
	Number of up, down, left, and right saccades	Number of occurrences
	SCL: Skin Conductance Level	Continuous
	Frequency of Skin Conductance Response (SCR) peaks	Number of occurrences
	Δ_{all}[Hb]: The Overall Brain Oxygenation Level	Continuous

The device offered a sensor pad comprising 18 optical sensors that record fNIRS signals with a sampling frequency of 20 Hz. We attached this sensor pad to the participant's forehead to monitor hemodynamic responses in the frontal lobe of the brain. During the recording session, we asked participants to refrain from moving their heads and the non-dominant hand to minimise the occurrence of motion artefacts.

3.3 Ground Truth

3.3.1 Pre-study Survey. For each of the four topics, we asked participants to rate their stance on the topic using an 11-point Likert scale (−5: *I agree with the con stance* to +5: *I agree with the pro stance*). We presented the pro and con stances according to Table 1. Participants also rated their interest in and familiarity with each topic using a 5-point Likert scale (1: *least interested* to 5: *completely interested*) and a 10-point Likert scale (1: *least familiar* to 10: *most familiar*), respectively.

3.3.2 Post-study Survey. After completing the data collection, we asked participants to reevaluate the stimuli they have seen in the study. Each participant rated the expressiveness of each stimulus on a 7-point Likert scale (1: *very weak* to 7: *very strong*). Each question was accompanied by the corresponding stimulus.

3.3.3 Participants' Ideological Stances. We gathered the users' ideological stances through the pre-study survey's responses. These were used to determine the congruence of stances between each participant and each stimulus. A stimulus S is considered *congruent* with participant P if the stances of S and P were in agreement. On the other hand, if the stances of S and P were opposite, S is *dissenting* with P.

We employed a threshold of 0 on the stance ratings (ranging from −5 to +5) to determine the participants' attitudes. For example, on the topic of climate change, a positive score implied the participant's stance aligned with the idea that climate change is man-made (i.e., "*I believe humans are primarily responsible for climate change.*"). Conversely, a negative score represented the stance that climate change is not man-made (i.e., "*I believe humans are not primarily responsible for climate change.*"). Participants who rated 0 on a topic were considered as having a neutral attitude on that particular topic. In our data analysis, we discarded any stimulus exposure that involved participants who had a neutral stance on a topic.

Among 33 participants who joined Study 1, we observed that most participants aligned themselves with the *pro* stances of every topic: progressivism (pro : con : neutral = 27 : 3 : 3), climate change (pro : con : neutral = 30 : 1 : 2), multiculturalism in Australia (pro : con : neutral = 30 : 1 : 2), and feminism (pro : con : neutral = 28 : 4 : 1).

4 STUDY 1: RESULTS

We analysed the data collected in Study 1 and examined the effects of ideological congruency on dwelling time, behavioural data, and physiological signals.

4.1 Dwelling Time

A one-way repeated measures ANOVA was performed on the amount of time each participant spent with each stimulus. We set the independent variable to be the congruence of ideological stance between the participant and the stimulus, which had two levels: congruent (C) and dissenting (D). For both text and image stimuli, we found that participants spent significantly more time with dissenting stimuli than congruent stimuli (text: $\mathcal{F}(1, 31) = 18.911, \eta_p^2 = 0.37, p < 0.001$; image: $\mathcal{F}(1, 29) = 4.416, \eta_p^2 = 0.13, p = 0.0444$). We found a weaker effect size (text: $\eta_p^2 = 0.37$, image $\eta_p^2 = 0.13$) for image stimuli.

4.2 Eye Tracking Measures

4.2.1 Preprocessing. We first obtained the raw gaze data, which consisted of the (x, y) coordinates on the projection screen. Subsequently, we used the Tobii Pro Lab's I-VT gaze filter [79] to estimate the velocity of the participant's eye movement. Those with a velocity below the threshold were considered *fixations* – a type of eye movement where the eyes are focused on one point. Those with a higher velocity were treated as *saccades* – rapid eye movement from one point to the other. For each stimulus, we obtained the eye-tracking features by calculating the number of fixations and the number of saccades in each of the four directions (up, down, left, and right) during the exposure to the stimulus.

4.2.2 Data Analysis. Since our eye tracking features consisted of count data which are often not normally distributed, we applied a Friedman test on the counts of fixations and saccades of both text stimuli data and image stimuli data. We subsequently corrected the p-values using a one-step Bonferroni correction.

For text stimuli, we found that participants exhibited significantly more fixations ($p_{corr} = 0.0174, \chi^2 = 12.461$), and right saccades ($p_{corr} = 0.0036, \chi^2 = 15.384$) with dissenting stimuli than congruent stimuli. For image stimuli, we observed significantly more fixations ($p_{corr} = 0.0283, \chi^2 = 11.560$) when viewing dissenting stimuli compared to congruent stimuli.

4.3 Electrodermal Activity

4.3.1 Signal Preprocessing. EDA signals recorded from a wearable device may contain motion artefacts. We, therefore, applied a lowpass filter with a cutoff frequency of 3 Hz to remove potential high-frequency motion artefacts. Subsequently, we applied a highpass filter with a 0.05 Hz cutoff frequency to extract the skin conductance responses (SCR) and the skin conductance level (SCL). SCR peaks were then identified by applying a peak detection algorithm to the SCR signals. Lastly, we derived two EDA measures: the mean of SCL and the count of SCR peaks throughout the period of exposure to a stimulus.

4.3.2 Data Analysis. Similar to the eye tracking data analysis, a Friedman test was performed on the EDA features and a one-step Bonferroni correction was used to correct the p-values. For text stimuli, we found that participants exhibited significantly greater counts of SCR peaks on dissenting statements ($p_{corr} = 0.0052, \chi^2 = 15.695$) than congruent statements. For image stimuli, however, we did not detect any significant effects of ideological congruence on EDA statistics.

4.4 Brain Hemodynamic Responses

4.4.1 Signal Preprocessing. The fNIRS we used in this study measured the optical density in two near-infrared frequencies, 730nm and 850nm. However, these signals are susceptible to noise, such as motion and physiological artefacts. Thus, for each participant, we first identified and discarded data that were distorted because of bad optode placement, i.e., when they were obstructed by hair or interfered with ambient light. Bad optode placement was considered if either (1) 90% of the optode's raw optical density fell outside an acceptable range of [400 mV, 4000 mV]; or (2) the raw optical density's coefficient of variation (defined as the ratio of the signal's standard deviation and mean) exceeded 20%.

Subsequently, we corrected noise and motion artefacts in the signals using 10-second time epochs. This involved two steps; first, we applied a bandpass filter with cutoff frequencies between [0.001 Hz, 1 Hz] on the optical densities to filter out signals from irrelevant frequency bands. Subsequently, the Temporal Derivative Distribution Repair (TDDR) algorithm [32] was applied to the filtered optical densities to correct motion and physiological artefacts.

We also manually removed parts of the recordings that consisted of suspected motion artefacts, i.e., rapid spikes in the signal which were caused by participants' body movement. We then subtracted the optical densities with the initial 5-second baseline. The baseline was recorded before the data collection started when participants were sitting still for about 20 seconds. The baselined optical densities were converted to oxygenated haemoglobin and deoxygenated haemoglobin concentrations ([HbO] and [HbR]) using the modified Beer-Lambert law [6]. [HbO] and [HbR] were then standardised within each participant to mitigate the effects of individual differences. Subsequently, we subtracted [HbR] from [HbO] and obtained the brain oxygenation level, Δ[Hb] = [HbO] − [HbR]. This step was done in order to improve the signal strength. Then, for each participant, we obtained the overall oxygenation level, Δ_{all}[Hb], by averaging Δ[Hb] across all available optodes. We opted to use Δ_{all}[Hb] to represent the overall changes in the forehead hemodynamic activity. Lastly, for each stimulus exposure, we calculated the mean Δ_{all}[Hb] for the exposure period.

4.4.2 Data Analysis. We applied a one-way repeated measures ANOVA on the mean overall oxygenation level, Δ_{all}[Hb], for each time window of stimulus exposure. However, we found no significant effects of ideological congruence on the mean overall oxygenation levels.

4.5 Summary and Lessons Learned

Our findings indicate that participants tended to spend more time and exhibited more fixations when facing ideologically dissenting stimuli. This implies that dissenting information might hinder or disrupt the comprehension process. However, it was inconclusive whether cognitive biases did contribute to this phenomenon. One possible assumption could be that ideologically dissenting text stimuli (i.e., the *con* statements) were more cognitively demanding than ideologically congruent stimuli [94]. Alternatively, since the make-up of the study participants were predominantly aligned with the *pro* statements, the dissenting stimuli may also have systematically caused longer dwelling time.

Although we found a significant effect on the counts of SCR during exposure to text stimuli, it remained inconclusive whether physiological signals are reliable indicators of cognitive biases. As a potential explanation, the study design may have introduced confounding factors: we did not provide a time gap between two consecutive stimuli, so-called inter-stimuli intervals (ISI). Due to the lack of ISI, the stimulus-related physiological responses may not reflect the reactions induced by the stimulus itself but those induced by the preceding stimuli.

In addition, we observed that image stimuli yielded less expressiveness than text stimuli for two reasons. First, we found no significant effect of the image stimuli's ideological congruence on dwelling time. Secondly, we found that the self-report expressiveness ratings on image stimuli were significantly lower than on text stimuli (one-way repeated measures ANOVA: $\mathcal{F}(1, 31) = 5.808, \eta_p^2 = 0.16, p = 0.022$).

Visual information is one of the most prevalent media on the Internet and is highly contextual, usually presented together with text information [89]. In contrast to text stimuli, image stimuli can thus be ambiguous, leading to different interpretations in different individuals.

To eliminate possible confounds, we conducted a follow-up study, which (1) ensured that the polarising statements successfully induced biased information processing, (2) used a reliable ground truth for the induced biased information processing, and (3) allowed us to observe clearer changes in physiological signals in response to each stimulus.

5 STUDY 2: DESIGN

As Study 1 was inconclusive as to whether cognitive biases were induced, we cannot draw any conclusions as to what extent physiological signals can be used to infer the presence of cognitive biases yet. We, therefore, designed and conducted Study 2 to address the same question as Study 1 – are physiological signals reliable, objective measures of cognitive biases? Study 2 comprised a similar approach in that we exposed participants to a series of polarising statements, but revised the experimental design to account for potential confounding factors.

5.1 Stimuli Selection

We employed 62 text stimuli in Study 2. We decided to expose our participants to a wider range of opinion statements. Thus, we aimed to increase the external validity of Study 2 by diversifying our stimuli and obtaining more observations. We extended the number of topics to eight in Study 2: *progressivism*, *climate change*, *feminism*, *multiculturalism in Australia*, *vegetarianism*, *renewable energy*, *abortion*, and *same-sex marriage*. We used the 32 original text stimuli from Study 1 and introduced 30 additional stimuli for the four new topics. We provide details of each new topic in the following paragraph. Informed by Study 1, we opted for not using image stimuli, as they proved difficult to limit confounding factors like the expressiveness and ambiguity of the images.

We included **vegetarianism** as one of the new topics because of an increasing debate (about 12% of Australians identify as vegetarians [101]) between proponents of vegetarianism (i.e., those who do not eat meat) and its opponents (i.e., those who support meat consumption). Meanwhile, we selected **renewable energy** since it is contextually parallel to the topic of climate change. In Australia, there has been a growing political debate between supporters and opponents of renewable energy [23]. We also selected **abortion** and **same-sex marriage** because they are part of the discussions on feminism and progressivism. Although abortion has been legalised in Australia, a notable proportion of pro-life messages still exist on the Internet [1]. Similarly, same-sex marriage in Australia was a heated debate during the 2017 marriage law survey [84]. While the poll showed that the majority of Australians (61%) expressed support for same-sex marriage, there was a significant proportion of those who voted "no" [33].

The 30 new stimuli were gathered from the Procon.org website. The ideological stances were counterbalanced, i.e., there was an equal number of pro and con statements. We controlled the length of each text stimulus to be around 50 to 80 words. In addition, we ensured that no text stimulus had a score lower than 30 according to the Flesch reading ease score [34], which is equivalent to the university level. In Table 3, we summarise the pro and con ideologies for each of the four additional topics.

Similar to Study 1, we presented the stimuli on a computer monitor. Each stimulus was displayed with the same font (Verdana 27 pt, in dark grey colour), double line spacing, centred-justified alignment, and white background. Figure 2 gives an example of the text stimuli in Study 2.

5.2 Study Protocols

5.2.1 Experimental Design. We conducted the study with a within-subject design with the independent variable being participants' congruent or dissenting opinion (i.e., two levels). Our dependent variables consisted of behavioural measures (dwelling time and eye tracking data), physiological measures (EDA and brain hemodynamic responses), and self-report measures (stimulus-wise ideological alignment, likelihood to share the stimulus, and cognitive effort). We present a list of study variables of Study 2 in Table 4.

5.2.2 Participants. We invited 31 participants (16 female, 13 male, and 4 preferred not to disclose) to Study 2. The mean age was 29.41 (SD= 11.17) and ranged from 18 to 68 years old. Of those who disclosed their age, 6 were between 18 and 20 years old, 10 were in their 20s, 6 were in their 30s, 2 were in their 40s, and 6 were 50 years old or older. All participants reported that they were either native, bilingual, or professional users of English. For their highest level of education, 10 had year 12 education, 1 had certificate III/IV education, 6 had a bachelor's degree, 4 had a graduate diploma/certificate, and 10 had a postgraduate degree. We excluded two participants for fNIRS data analysis and one participant for EDA data analysis since their recordings were mostly corrupted or missing.

5.2.3 Procedure. Similar to Study 1, Study 2 took place in a quiet room where participants were seated in a comfortable position in front of a 24-inch monitor. We first informed each participant about the purpose and protocols of the study. After receiving their written consent, we asked the participants to answer a pre-study survey and calibrated the physiological sensors. After that, participants went through a warm-up round to familiarise themselves with the protocols. We presented participants with a series of four text stimuli on the topic "*Should zoos exist*?". These warm-up stimuli were sourced from Procon.org[7].

In this study, we exposed participants to stimuli differently from Study 1. For each stimulus, participants first read the stimulus statement. Once they finished reading it, they responded to an in-study survey, which asked participants three questions regarding the stimulus. After providing their responses, participants entered a 15-second resting period, where we asked them to close their eyes and count from 1 to 15. A 15-second timer was placed on a screen. Once the timer counted down to 0, participants proceeded to the next block by clicking on the "next" button. The presentation order of the stimuli was randomised.

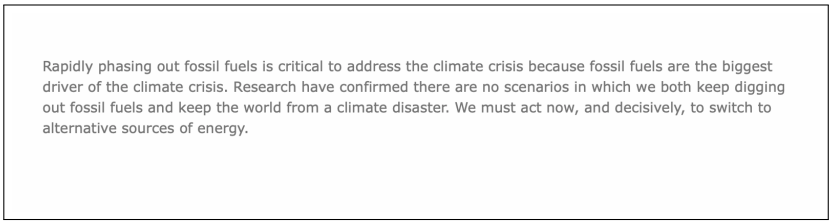
We introduced a 15-second resting period as an inter-stimulus interval (ISI) in order to observe clearer physiological changes. We decided that 15 seconds would be an appropriate ISI since it allowed sufficient time to observe hemodynamic responses, which typically take three to five seconds to reach a peak and a few seconds to decay [69, 107].

After participants finished the warm-up round, they entered the data collection round. In this round, we presented participants with a series of 62 text stimuli from the eight topics mentioned. Like the warm-up round, participants read the stimulus, responded to an in-study survey, and entered a 15-second resting period. Each stimulus was presented in a randomised order; each stimulus' topic and stance were also randomised. Upon completion, we engaged

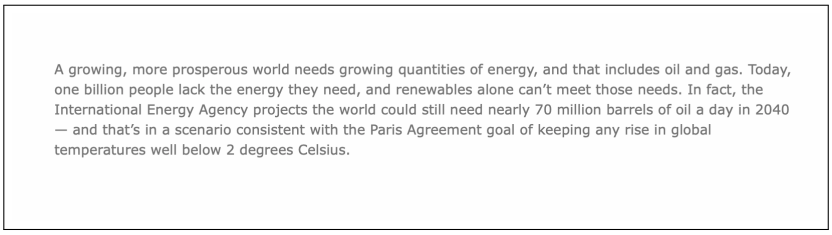
[7] https://www.procon.org/headlines/zoos-top-3-pros-and-cons/

Table 3: Additional topics in Study 2 and their ideological ends

Topic	Pro stance	Con stance
Vegetarianism	I support vegetarianism and oppose meat consumption	I support meat consumption and oppose vegetarianism
Renewable Energy	I believe renewable energy is necessary	I believe renewable energy is not necessary
Abortion	I think abortion should be legal	I think abortion should be prohibited
Same-sex marriage	I think same-sex marriage should be legal	I think same-sex marriage should be prohibited



(a) A pro stance on renewable energy



(b) A con stance on renewable energy

Figure 2: Examples of stimuli presentation for Study 2

participants for a brief interview and compensated them with a $20 cash voucher. The study took approximately 90 minutes.

5.2.4 Sensors. Throughout the experiment, we recorded physiological data from the participants. In a similar fashion to Study 1, we employed the Empatica E4 wristband to record EDA and the BIOPAC fNIR Sensors 2000 to record brain oxygenation levels from the forehead. We used a Tobii Pro Nano[8] to record the participant's eye-tracking data with a sampling frequency of 120 Hz. We asked participants to refrain from moving their heads throughout the data collection period to prevent the occurrence of motion artefacts.

5.3 Ground Truth

5.3.1 Pre-study Survey. For each of the eight topics, we asked participants to rate their stance on the topic on a continuous slider scale of 0 (*I agree with the **con** stance*) to 100 (*I agree with the **pro** stance*). Unlike Study 1, we used a continuous scale for ideological stance since it provides more granularity for assessing the participants' stances on a spectrum. In addition, we asked participants to rate their interest and familiarity with each topic on a scale from 1 (*least interested/familiar*) to 5 for (*most interested/familiar*).

5.3.2 In-study Survey. For each stimulus, we asked participants to report the congruence of ideological stance between them and the statement, the likelihood to share it on their social media, and the cognitive effort spent reading it, by asking three questions: (Q1) *How much does the statement align with your beliefs?*; (Q2) *How likely are you to share this statement on your social media?*; and (Q3) *How much effort did you put into reading this statement?*. Participants gave their ratings using a 5-point Likert scale (1: *least aligning/likely/effortful* to 5: *most aligning/likely/effortful*). The in-study survey was triggered each time participants finished reading a stimulus.

5.3.3 Participant-stimulus Ideological Stance. On each topic, we determined the ideological alignment of each participant from their self-reported stance (from 0 to 100) in the pre-study survey. Similar

[8]https://www.tobiipro.com/product-listing/nano/

Table 4: Summary of Independent and Dependent Variables in Study 2

Variables	Measures	Scale
Independent Variables	Participant-Stimulus Ideological Congruency	2 levels (congruent and dissenting)
	Topic Interest	2 levels (high and low)
	Topic Familiarity	2 levels (high and low)
Dependent Variables	**Behavioural**	
	- Dwelling Time	Continuous
	- Number of Fixations	Number of occurrences
	- Number of up, down, left, and right saccades	Number of occurrences
	Self-report	
	- Q1: participant-stimulus ideological congruence	5-Likert scale
	- Q2: likelihood to share the stimulus on one's social media	5-Likert scale
	- Q3: effort spent reading the stimulus	5-Likert scale
	Physiological	
	(Time windows {2.5s, 5s, 10s} × {EXP1, EXP2, POST})	
	- SCL: Skin Conductance Level	Continuous
	- Frequency of Skin Conductance Response (SCR) peaks	Number of occurrences
	- Δ_{all}[Hb]: The Overall Brain Oxygenation Level	Continuous

to Study 1, we applied a threshold of 50 on the stance ratings. A rating of more than 50 represented an ideological stance that supports the pro stance. Conversely, ratings less than 50 were considered to support the con stance.

Using the abovementioned thresholds, our 31 participants identified their stances as follows: climate change (pro : con : neutral = 28 : 1 : 2), feminism (pro : con : neutral = 25 : 2 : 4), progressivism (pro : con : neutral = 24 : 5 : 2), multiculturalism in Australia (pro : con : neutral = 30 : 0 : 1), vegetarianism (pro : con : neutral = 12 : 12 : 6), renewable energy (pro : con : neutral = 29 : 1 : 1), same-sex marriage (pro : con : neutral = 26 : 4 : 1), and abortion (pro : con : neutral = 27 : 3 : 1).

Subsequently, we defined a score that describes the ideological congruency between the participant and the stimulus. The score was in a range between −50 (*the participant's stance is completely opposite of the stimulus*) and +50 (*the participant's stance completely aligns with the stimulus*). A positive score implied that the stances of the participant and the stimulus were in the same direction, and vice versa. The congruency score between participant p and stimulus s, $Congruence(p, s)$, can be derived by applying formula 1. We denoted $Pos(s)$ as the ideological stance of the stimulus s, which took a binary value of +1 if the stance was aligned with the pro opinion or −1 if the stance was aligned with the con opinion. $Stance(p)$ is the self-report stance of the participant p, ranging from 0 to 100.

$$Congruence(p, s) = (Stance(p) - 50) \times Pos(s) \quad (1)$$

For example, if a person rated themself with 80 out of 100 on the topic of abortion and a stimulus stated an anti-abortion statement, the congruency score between them would be $(80 - 50) \times (-1) = -30$.

In this study, we considered stimulus exposures with a congruency score greater than +20 and those with a score lower than −20 to be ideologically *congruent* (C) and *dissenting* (D) respectively. We discarded data points where the congruency score was between −20 and +20 as they were considered neutral or weak in inclination. The scale for the score was continuous with the mean of $M = 0$ and $SD = 37.32$.

6 STUDY 2: RESULTS

We analysed the effects of the congruency between the stimuli's ideologies and the participants' leanings on behavioural, physiological, and interaction measures collected during the study. The goal was to examine physiological expressions of cognitive biases that may be experienced when aligning with or distancing oneself from content items. In the following, we describe our analysis and findings along with the training of a classifier to detect the participant-stimulus ideological congruency from interaction and physiological data on whether participants encountered attitudinal information.

6.1 Effects of Ideological Congruency

We applied a one-way repeated measures ANOVA on the amount of time each participant spent with each stimulus. Similar to Study 1, the independent variable was the congruence of ideological stance between the participant and the stimulus, which had two levels: congruent (C) and dissenting (D). Accordingly, we examined the effects of opinion congruency on dwelling time, self-report measures,

eye tracking measures, and physiological (electrodermal activity and hemodynamic responses). Table 5 reports statistical results of the self-report and behavioural measures. Table 6 reports the statistical results of the physiological measures.

6.1.1 Behavioural Measures. We found that participants spent significantly more time with ideologically dissenting stimuli than with congruent stimuli (C: 12.35 ± 7.46 seconds, D: 12.92 ± 7.26 seconds, $\mathcal{F}(1, 30) = 5.713, \eta_p^2 = 0.160, p = 0.023$)

For eye tracking measures, similar to Study 1, throughout the period of stimulus exposure, we calculated the number of fixations and the number of saccades in each of the four directions: up, down, left, and right. Since the length of each exposure was not identical, we normalised the measures by dividing each of them by the stimulus dwelling time. We found no significant effect of opinion congruency on the normalised eye-tracking measures.

6.1.2 Self-report Measures. We performed a similar analysis on the self-reported ratings for each stimulus. We examined (Q1) the ideological congruence between the participant and the stimulus, (Q2) participants' likelihood to share the stimulus on their social media, and (Q3) their cognitive effort spent reading it.

We found that Q1 and Q2 responses from congruent stimuli were significantly higher than those from dissenting stimuli (Q1: C: 3.90± 0.92, D: 2.29 ± 1.12, $\mathcal{F}(1, 30) = 203.481, \eta_p^2 = 0.871, p < 0.001$; Q2: C: 1.83 ± 1.03, D: 1.23 ± 0.51, $\mathcal{F}(1, 30) = 51.564, \eta_p^2 = 0.632, p < 0.001$). This confirms the internal validity of the stimulus materials as the participants' general tendency toward a topic ($Congruence(p, s)$) and their content-specific alignment (Q1) were congruent. Specifically, we found that $Congruence(p, s)$ and Q1 were strongly correlated (Pearson $r = 0.737, p < 0.001$). Moreover, participants with general tendencies in favour of a topic were more willing to share content that aligned with their views. Q3 responses for congruent and dissenting stimuli were not significantly different from each other (Q3: C:2.98 ± 1.20, D:2.76 ± 1.23, $\mathcal{F}(1, 30) = 3.375, \eta_p^2 = 0.101$, n.s.).

6.1.3 Physiological Measures. The task design in Study 2 allowed us to observe physiological changes both during and after stimulus exposure. Thus, we analysed the collected physiological signals in three different time windows: a period during the beginning of stimulus exposure (EXP1: the first 0 to w seconds), a period during the end of stimulus exposure (EXP2: the final w seconds), and a period after stimulus exposure (POST: the first 0 to w seconds after exposure). We analysed the data using three different window sizes (w): 2.5, 5, and 10 seconds. The choices of window size followed those commonly used in prior EDA [11, 22] and fNIRS studies [2, 40]. To ensure that our window analysis is valid, we discarded any exposure that lasted shorter than the defined window size.

Additionally, we corrected the temporal drift in SCL by subtracting the SCL values in the baseline window from the SCL values in the analysis window. For each stimulus, we used the final 2 seconds of the resting period (i.e., the ISI) before the participant started reading it as the baseline window.

We followed the same signal preprocessing pipeline as in Study 1. We examined 3 × 3 dependent variables. From EDA data, we calculated the mean of SCL and the frequency of SCR. For hemodynamic responses, we obtained the mean overall oxygenation levels, Δ_{all}[Hb]. Each of these measures was calculated in the three time windows: EXP1, EXP2, and POST.

In a similar manner, we ran a repeated measures ANOVA on each of the dependent variables. We did not detect a significant effect of opinion congruency on the mean of SCL in any time window. However, we observed a trend during the first 10 seconds of stimulus exposure (EXP1 period) that the mean SCL was higher when presented with dissenting stimuli (C: 0.000436 ± 0.143, D: 0.0203 ± 0.188, $\mathcal{F}(1, 29) = 4.242, \eta_p^2 = 0.127, p = 0.0502$).

We did not find a significant effect of opinion congruency on the overall oxygenation levels; yet, we found significant effects on the overall oxygenation levels of the subgroup of participants who reported low interest in a topic. We discuss this finding in detail in the following section.

6.2 Effects of Interest and Familiarity

We examined whether participants' interest in and familiarity with a topic influenced their self-report, behavioural, and physiological expressions. To do so, we considered subgroups of participants with high/moderate/low interest and familiarity with a topic.

For each topic, we set a threshold of 3 on the interest (1-5) ratings. We considered those who rated topic interest as 4 or 5 to have *high interest*. Participants who rated 3 were regarded as having *moderate interest*. Lastly, those who rated 1 or 2 on interest were deemed as *low interest*. We also applied the same threshold on the familiarity (1-5) ratings to form participant groups with high, moderate, and low familiarity.

There were a total of 1482 observations across 31 participants. When filtered by topic interest, there were 232 observations across 19 participants in the low-interest group and 900 observations across 29 participants in the high-interest group. When filtered by topic familiarity, there were 354 observations across 18 participants in the low-familiarity group and 522 observations across 25 participants in the high-familiarity group.

We analysed the high-interest and low-interest groups separately by employing a one-way repeat measures ANOVA on each of the measures. While we detected some effects of topic interest, we found no effect from familiarity; thus, we provide the analysis only for topic interest in the following. Table 5 gives the testing results, including the sample size for each subgroup.

6.2.1 Dwelling Time. In line with the general results, we found that participants with a higher interest in a topic spent significantly more time with dissenting information (C:11.96 ± 6.46, D: 13.20 ± 7.49, $\mathcal{F}(1, 28) = 13.470, \eta_p^2 = 0.829, p < 0.001$). We also detected a greater effect size in the high-interest group (compared to the general group, $\eta_p^2 = 0.160$), which indicated that the effect of ideological congruency was stronger in participants with high interest. Meanwhile, we found no significant effect when considering data from low-interest individuals.

6.2.2 Self-report Measures. We obtained consistent results from both subgroups: they tended to rate both Q1 and Q2 higher for congruent stimuli. We observed a greater effect size on Q2 in the high-interest group (high-interest group: $\eta_p^2 = 0.655$, general group: $\eta_p^2 = 0.632$), indicating that individuals were more likely to share attitude-confirming contents as they were more interested in the

topic. In addition, we found that participants with low interest reported significantly higher effort (Q3) when reading congruent statements than dissenting ones (C: 2.98 ± 1.20, D:2.76 ± 1.23, $\mathcal{F}(1, 18) = 9.348, \eta_p^2 = 0.341, p = 0.006$).

6.2.3 Physiological Measures. When examining data from low-interest individuals, we detected significant effects of ideological congruency on the overall oxygenation levels during the EXP1 period. Our analysis showed that the effects were significant in window lengths of 2.5 and 5 seconds, where participants tended to exhibit higher oxygenation levels when facing dissenting information (2.5-second window: C: -0.18 ± 1.08, D: 0.10 ± 1.11, $\mathcal{F}(1, 16) = 5.352, \eta_p^2 = 0.250, p = 0.034$; 5-second window: C: -0.16 ± 0.99, D: 0.059 ± 1.07, $\mathcal{F}(1, 16) = 4.607, \eta_p^2 = 0.223, p = 0.048$). As higher oxygenation levels associate with more neural activation, our results suggested that ideologically diverging information induced higher neural activity than congruent information. We found no significant effect when considering high-interest individuals.

6.3 Building a Bias Classifier

To examine our measures as indicators of cognitive biases, we performed a binary classification on the collected data to detect and distinguish the exposure to ideologically congruent stimuli (C) from ideologically dissenting (D) ones. We extracted the input features of the classifiers from statistical values of the EDA and brain oxygenation levels. Each of the features was extracted in a 2-second time window in each observation period (EXP1, EXP2, and POST). Statistics include the mean, standard deviation, median, kurtosis, skewness, and slope. We also included eye-tracking features, which were the counts of fixations and saccades in different directions (up, down, left, and right). Due to counterbalancing in the study design, our dataset (4960 samples) was perfectly balanced between the congruent and dissenting conditions (class ratio C : D = 2480 : 2484).

We trained a model using the following classifiers: linear discriminant analysis (LDA), support vector machine (SVM) with an RBF kernel, random forest [13], and XGBoost [36]. We evaluated the models by using the average accuracy across a 5-fold cross-validation. The average accuracy was calculated from the mean of the validation accuracy for each fold. For tree-based models, we performed hyperparameter tuning using a randomised search for the number of trees and the maximum depth. The optimal parameters were 1600 trees and 30 levels for random forest, and 1500 trees and 6 levels for XGBoost.

We found that the highest accuracy achieved was 55.27% on average through the XGBoost algorithm. The result, however, indicated that our classifier performed barely above the performance of a ZeroR classifier, i.e., the level of chance (50.04% accuracy for our dataset). To ensure that the model performance scores were not obtained by chance, we performed a permutation test [78] on each of the classification algorithms. We found that all models except ZeroR achieved a p-value lower than 0.05, indicating that the employed models can give better predictions than the chance level with 95% confidence. Table 7 summarises each model's classification performance and the p-value of the permutation test.

7 DISCUSSION

To avoid information overload and effectively categorise the vast information available online, people often resort to mental shortcuts to make quick judgments about new information. These shortcuts can lead to a biased interpretation of that information and hence form what is called cognitive biases [102]. In the presented studies, we explored the indicators of cognitive biases in information consumption in two experiments, in which we exposed participants to ideologically polarising stimuli while collecting self-reports, behavioural, and physiological measures. Study 1 showed that some of our results were inconclusive in terms of physiological measures due to a lack of time gaps between subsequent stimuli. Hence, we were unable to isolate the effect of the stimuli on participants' opinion-related reactions. However, we found that participants spent more time with ideologically dissenting information but it was unclear whether this was due to the influence of their biased perception of the topic or whether some stimuli were more cognitively demanding in the way they were presented than others.

In Study 2, we addressed this limitation by redesigning the study and collecting not only behavioural (dwelling time and eye tracking) and physiological measures (EDA and hemodynamic responses) but also self-reports on topic interest and familiarity as they have been shown to influence the depth of information processing [63]. We ensured internal validity as participants demonstrated that their general tendency on a topic and their content-specific alignment (Q1) were consistent. Secondly, we introduced the use of inter-stimuli intervals (ISI) in Study 2. This allowed us to observe clearer physiological responses following stimulus exposure. Thirdly, we exposed participants to a greater range of opinion statements, thus increasing the external validity of the study.

In the remainder of this section, we discuss the outcomes of both studies focusing on the behavioural and physiological expressions of cognitive biases when viewing different opinions. We first discuss the effects of ideological congruency on dwelling time found in both studies. Subsequently, with Study 2 suggesting that topic interest influences the effects of ideological congruency, we discuss topic interest as a factor of biases. Lastly, we discuss the implications of building bias-aware systems, their feasibility, potential impact, as well as some ethical considerations.

7.1 Behavioural Expressions of Biases

In both studies, we observed that participants tended to spend more time with dissenting than opinion-confirming information. As in Study 2, the effects became stronger when considering participants with high interest in a specific topic. Our results support prior findings on selective exposure [37, 100]. Meanwhile, the results draw contrast to some prior works [68, 91], which stated that people tend to spend more time viewing confirmatory information.

Research on selective exposure has produced mixed results in terms of behavioural measures. With our studies showing different results from some of the existing literature, study designs may influence the behaviour of the participants and thus their behavioural expression of biases. Our study exposed participants to discrete pieces of information – i.e., participants read the stimulus contents one by one. Research by Garrett [37] and Taber and Lodge [100]

Table 5: Inferential statistics of Study 2's stimulus dwelling time and self-report measures. N and n denote the number of included participants and the number of included stimulus exposure, respectively. We denote **, ***, and **** for significance levels of 0.05, 0.01, and 0.001, respectively.

Measure Sample sizes	**General** ($N = 31, n = 1482$)	**Low interest** ($N = 19, n = 232$)	**High interest** ($N = 29, n = 900$)
Dwelling time	**D > C**** $\mathcal{F}(1, 30) = 5.713$ $\eta_p^2 = 0.160, p = 0.023$	n.s. $\mathcal{F}(1, 18) = 2.560$ $\eta_p^2 = 0.126, p = 0.0995$	**D > C***** $\mathcal{F}(1, 28) = 13.470$ $\eta_p^2 = 0.324, p = 0.001$
Q1	**C > D****** $\mathcal{F}(1, 30) = 203.481$ $\eta_p^2 = 0.871, p < 0.001$	**C > D****** $\mathcal{F}(1, 18) = 37.064$ $\eta_p^2 = 0.673, p < 0.001$	**C > D****** $\mathcal{F}(1, 28) = 135.994$ $\eta_p^2 = 0.829, p < 0.001$
Q2	**C > D****** $\mathcal{F}(1, 30) = 51.564$ $\eta_p^2 = 0.632, p < 0.001$	**C > D***** $\mathcal{F}(1, 18) = 11.628$ $\eta_p^2 = 0.392, p = 0.003$	**C > D****** $\mathcal{F}(1, 28) = 53.279$ $\eta_p^2 = 0.655, p < 0.001$
Q3	n.s. $\mathcal{F}(1, 30) = 3.375$ $\eta_p^2 = 0.101, p = 0.076$	**C > D***** $\mathcal{F}(1, 18) = 9.348$ $\eta_p^2 = 0.341, p = 0.006$	n.s. $\mathcal{F}(1, 28) = 3.493$ $\eta_p^2 = 0.0792, p = 0.131$

Table 6: Inferential statistics of Study 2's physiological measures. N, n, and w denote the number of included participants, the count of included stimulus exposure, and the window size, respectively. We denote **, ***, and **** for significance levels of 0.05, 0.01, and 0.001, respectively.

Measure	**General**	**Low interest**	**High interest**
SCL during EXP1 ($w = 10$ seconds)	n.s. ($N = 30, n = 754$) $\mathcal{F}(1, 29) = 4.242$ $\eta_p^2 = 0.127, p = 0.0502$	n.s. ($N = 18, n = 112$) $\mathcal{F}(1, 12) = 0.921$ $\eta_p^2 = 0.0713, p = 0.356$	n.s. ($N = 28, n = 448$) $\mathcal{F}(1, 26) = 0.987$ $\eta_p^2 = 0.0365, p = 0.329$
Δ_{all}[Hb] during EXP1 ($w = 2.5$ seconds)	n.s. ($N = 29, n = 1397$) $\mathcal{F}(1, 27) = 1.133$ $\eta_p^2 = 0.0402, p = 0.296$	**D > C**** (**N = 17, n = 200**) $\mathcal{F}(1, 16) = 5.352$ $\eta_p^2 = 0.250, p = 0.034$	n.s. ($N = 27, n = 814$) $\mathcal{F}(1, 25) = 1.390$ $\eta_p^2 = 0.0526, p = 0.249$
Δ_{all}[Hb] during EXP1 ($w = 5$ seconds)	n.s. ($N = 29, n = 1228$) $\mathcal{F}(1, 27) = 1.865$ $\eta_p^2 = 0.0646, p = 0.183$	**D > C**** (**N = 17, n = 185**) $\mathcal{F}(1, 16) = 4.607$ $\eta_p^2 = 0.223, p = 0.048$	n.s. ($N = 27, n = 777$) $\mathcal{F}(1, 25) = 2.605$ $\eta_p^2 = 0.0943, p = 0.119$

Table 7: The Evaluation Scores for Bias Classification.

	ZeroR	**LDA**	**SVM**	**Random Forest**	**XGBoost**
Mean Accuracy (SD)	50.04 (0)	50.96 (0.02)	50.20 (0.001)	54.39 (1.94)	55.27 (2.74)
p-value	1.00	0.047	0.047	0.047	0.047

followed a similar protocol to our studies and produced congruent results with our work. On the other hand, works by Marquart [68], for example, comprised a different study design where the participants freely navigated information on the screen while their dwelling time was tracked through area-specific fixation time.

Regarding reading effort (Q3), we find that individuals tended to spend more time but reported less effort reading information with dissenting stimuli. Our results align with the theory of epistemic monitoring by Richter [86, 87], which states that ideological dissenting information disrupts the fluency of information processing. As a result, individuals economise their cognitive resources by allocating them to attitude-consistent information. The theory may explain our findings that the prolonged reading time for dissenting statements resulted from the participant's reduced fluency in comprehending inconsistent information. Subsequently, less reading effort implies that individuals tend to save up their cognitive resources to process congruent information.

7.2 Physiological Expressions of Biases

We found that topic interest influenced the effects of opinion congruency on physiological responses. When considering individuals with low interest in a topic, we detected significant effects on the brain oxygenation levels during the start of the stimulus exposure. Our findings indicate that individuals tended to exhibit higher neural activation levels when processing ideologically dissenting information. This result is in line with prior research on cognitive dissonance [12, 49, 110], suggesting higher neural activation when facing attitude-challenging information.

Our results add to the existing literature on psychophysiology. To the best of our knowledge, this is the first study to obtain these findings using fNIRS sensors in the context of information exposure. While the physiological research on information consumption has been limited, it will be interesting to devise future studies that observe the interactions between individuals' involvement with a topic and their ideological tendency through more objective measures like physiological data.

In addition, we detected a trend that the skin conductance levels (SCL) were higher in dissenting stimuli. However, the result remained statistically inconclusive. Our results drew parallels to a study by Ploger et al. [82] which investigated cognitive dissonance through video media consumption. Similarly, albeit not statistically significant, Ploger et al. found that SCL tended to be higher when facing attitude-challenging information. We argue that our attitude-dissenting stimuli may induce dissonance arousal [111] – i.e., the physiological by-product of cognitive dissonance. Nonetheless, future research may focus on the potential of EDA in detecting the psycho-physiological effects of ideologically polarising information.

7.3 Topic Interest as a Factor of Bias

By varying the analysis on subgroups of high and low-interest individuals, we found that topic interest impacted the occurrence of cognitive biases. In sum, higher topic interest strengthened the effects of ideological congruency on dwelling time and the likelihood of sharing the stimulus content. Lower topic interest, on the other hand, positively influenced the effects of ideological congruency on the reading effort (Q3) and the physiological measures (skin conductance levels and brain oxygenation levels).

Our finding is in line with prior research on selective exposure [29, 53, 90, 95], which states that topic interest is one of the influencing factors for the selective exposure effect. Our result is also supported by the two-step model of processing conflicting information by Richter [86, 87]. The theory states that people tend to use the perceived plausibility of the information as a heuristic: they tend to save up their cognitive resources on attitude-consistent information and process the information based on their beliefs. On the other hand, individuals with relevant background knowledge tend to process it in an informed and balanced way.

Interestingly, we did not find significant effects of topic familiarity on the occurrence of biases. Instead, we detected such effects from topic *interest*. Since we did not explicitly assess prior knowledge, future studies should consider the effects of topic knowledge and familiarity on bias occurrence.

7.4 Towards Bias-Aware Systems

The studies presented are a first step to building *bias-aware systems*, i.e., computing systems that detect and take into account the presence of cognitive biases in users [24]. The notion of bias-aware systems parallels cognition-aware systems coined by Bulling and Zander [15] as they pick up and adjust to cognitive states but with a focus on biases and predispositions. Our results feed into system frameworks, such as Nussbaumer et al. [77], which collect user-system interaction data, learn to detect cognitive biases from such data, and help users reduce their biases by providing feedback from bias detection.

With multimodal data collected in our study, we employed a range of machine learning algorithms on the collected data to classify exposures that involved congruent information from those with dissenting information. As our models barely outperformed chance, the challenge remains to build a bias-aware system based on a well-performing classifier.

Our studies, however, show some promising results in connecting physiological and interaction data with users' innate opinions and attitudes. For the field of human-computer interaction, identifying these markers and designing experiments around eliciting and measuring cognitive biases is the first step towards researching and building bias-aware systems. In the context of recent societal impacts of computing systems, we envision more research in and broader use of measuring tools for the presence and effects of user biases in the evaluation of computing systems. To this end, we also release our study materials, including content and data collection apparatus as supplementary materials.

Being able to quantify the occurrence and the effect of cognitive biases will allow researchers to closely study the influence that user interfaces and algorithms have on opinion formation. Systems capable of identifying biases will subsequently enable work to address and mitigate their effects. Hence, interventions can be designed and tested to help users overcome cognitive fallacies as a result of their biases and encourage them to engage more critically with computing systems and information. In the current climate of misinformation, sensing users' attitudes and reactions towards potentially biased information can help design better information

diets that help users break out of their filter bubbles, leave their silos of selected exposure and engage on a broader spectrum of ideas and opinions. Critical thinking and informed decision-making are critical for a healthy and diverse public discourse and have the potential to curtail the misinformation pandemic [60].

Finally, we would like to acknowledge the potential ethical implications of systems that sense biases, attitudes, and opinions. What can be used to identify and mitigate biases might as well be abused to reaffirm and steer people's beliefs, spread propaganda, and influence decision-making. The case of Cambridge Analytica has prominently demonstrated how people's attitudes can be derived from interaction data on social media platforms and used to influence opinion making [17]. Our research contributes to the systematic study of biases in the hopes that future work focuses on the demystification of how biases occur and what exacerbates or mitigates them.

8 LIMITATIONS

Despite Study 2 having addressed the main limitations of Study 1, there are a number of limitations we would like to discuss with regard to our study design and the interpretation of our findings.

First of all, our study design did not impose time constraints on each stimulus. Participants were free to spend as much time as they wanted with the stimulus until they clicked the next button. While this protocol allowed users to fully comprehend the stimuli materials, it introduced a number of limitations to the data analysis. The varying stimulus exposure time made it difficult to anticipate the temporal location of physiological reactions (i.e., the rise of oxygenation and skin conductance levels) regarding the stimuli. In addition, it was unclear how the reading motives of each participant affected their decision to end the stimulus exposure. Some participants may have tried to fully comprehend the material before clicking *next*, while others may have clicked next once they felt it unnecessary to further read the statement. We, however, took this into account by discarding data samples that lasted shorter than the analysis window size to mitigate the effects of shortened exposure.

Second, we were unable to assess the degree or strength of each stimulus's ideological tendency. As the primary source of our study stimuli, *ProCon.org* provides a collection of supporting and refuting information, which consists of opinions on different spectrums of attitude strength. In addition, our participants may perceive each stimulus individually, and therefore differently. During our post-study interviews, some participants reported they found some statements were not aligned with any particular ideological standpoint. The use of topic interest and familiarity as a subjective measure, however, helped us refine our analysis and isolate those cases where stronger tendencies may have been present.

The make-up of our study participants was also rather imbalanced in terms of ideologies. We found that most participants identified themselves as progressive, left-leaning, i.e., they mostly positioned themselves with the pro stances: they believed in man-made climate change and supported same-sex marriage. Since most participants were recruited from the university community, it was difficult for us to find people from conservative or right-leaning groups. While we tried to ensure that participants were exposed to an even number of congruent and dissenting stimuli, further studies with people with strong convictions and rather conservative and right-leaning attitudes are needed. For example, Knobloch-Westerwick et al. [55] found a greater leaning toward the US Republican party increased confirmation bias hinting that ideological alignment may have an effect on the creation and experience of biases.

In terms of physiological sensors, we used the Empatica E4 wristband to measure EDA. While the device is compact and unobtrusive, recent research has expressed concerns that E4-generated EDA signals are prone to motion artefacts and measurement noise and may not be as reliable as laboratory-grade devices [4, 10, 73]. Moreover, we only collected the hemodynamic responses using fNIRS from the forehead region. Although the placement of the sensor pad allowed unobtrusive data collection, it limited our observations to neural activities beyond the forehead area that is not covered by hair. Future works may investigate the hemodynamic activity from full-head fNIRS.

We used self-report ratings as pre-study questions to gauge participants' ideological inclination on, interest in, and familiarity with each topic, as well as in-study surveys to improve the internal validity of our study. While self-reports are convenient tools to collect information, they can be confounded by a range of factors, notably, memory, self-presentation [96] and preference falsification [59]. Moreover, written questions are susceptible to misinterpretation. For instance, some participants reported that they understood the question "*How much effort did you put in reading this statement*" (Q3) as the amount of cognitive resources spent on reading the statement, while some reported that they replied to Q3 by providing the degree of how well they understood the statement.

Moreover, we only used single-item questions to represent each of the self-report measures. This limits the external validity of our study since we did not use standardised, established subjective measures. For example, we did not examine the reading effort (Q3) using a well-established NASA-TLX questionnaire out of concern for participants' time and fatigue levels. Moreover, we did not assess the participant's general ideological alignment using, for instance, the Wilson-Patterson conservatism scales or performing an implicit association test [24]. Such tools may allow future research to uncover a more nuanced relationship between the strength of ideological conviction and individuals' experience of biases.

9 CONCLUSION

Biases as cognitive shortcuts help people cope with the vast amounts of online information but can also trap us in one-sided exposure and filter bubbles that reaffirm our existing beliefs. To study the occurrence and effects of biases on information consumption behaviour and decision-making, we set out to explore indicators for the occurrence of biases in physiological and interaction data. In this paper, we presented two experiments, in which we exposed users to opinionated statements on polarising topics while collecting physiological, behavioural, and interaction data. Our stimuli samples stated opinions that were either congruent or dissenting with participants' attitudes. We found that participants tended to generally spend more time processing statements that were incongruent with their own opinion. We further observed higher neural activity as indicated by certain brain regions' blood oxygenation

levels when participants were facing ideologically dissenting attitudes while having expressed relatively low interest in that topic. Our results demonstrate the existence of behavioural and physiological differences in the expression of congruency between people's innate opinions and ideologically tainted information, a first step towards building classifiers to detect cognitive biases.

Our study design and findings pave the way for future research in understanding the occurrence of cognitive biases with the goal of detecting them and quantifying their effects. The ability to equip systems with bias-awareness allows HCI researchers to study the role that design, algorithms, and content elements play in mitigating or exacerbating user biases.

ACKNOWLEDGMENTS

We thank the participants of our studies.

REFERENCES

[1] Robert Ackland and Ann Evans. 2017. Using the web to examine the evolution of the abortion debate in Australia, 2005-2015. In *The Web as History*, Niels Brügger and Ralph Schroeder (Eds.). University College London, Gower Street, London WC1E 6BT, 159–189.

[2] Haleh Aghajani, Marc Garbey, and Ahmet Omurtag. 2017. Measuring Mental Workload with EEG+fNIRS. *Frontiers in Human Neuroscience* 11 (2017). https://doi.org/10.3389/fnhum.2017.00359

[3] Faisal Alatawi, Lu Cheng, Anique Tahir, Mansooreh Karami, Bohan Jiang, Tyler Black, and Huan Liu. 2021. A Survey on Echo Chambers on Social Media: Description, Detection and Mitigation. *arXiv preprint arXiv:2112.05084* (2021). https://arxiv.org/abs/2112.05084

[4] Ebrahim Babaei, Benjamin Tag, Tilman Dingler, and Eduardo Velloso. 2021. A Critique of Electrodermal Activity Practices at CHI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 177, 14 pages. https://doi.org/10.1145/3411764.3445370

[5] Ricardo Baeza-Yates. 2018. Bias on the Web. *Commun. ACM* 61, 6 (May 2018), 54–61. https://doi.org/10.1145/3209581

[6] Wesley B. Baker, Ashwin B. Parthasarathy, David R. Busch, Rickson C. Mesquita, Joel H. Greenberg, and A. G. Yodh. 2014. Modified Beer-Lambert law for blood flow. *Biomed. Opt. Express* 5, 11 (Nov 2014), 4053–4075. https://doi.org/10.1364/BOE.5.004053

[7] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132. https://doi.org/10.1126/science.aaa1160 arXiv:https://www.science.org/doi/pdf/10.1126/science.aaa1160

[8] Mathias Benedek and Christian Kaernbach. 2010. A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods* 190, 1 (2010), 80–91. https://doi.org/10.1016/j.jneumeth.2010.04.028

[9] Alexander Benlian. 2015. Web Personalization Cues and Their Differential Effects on User Assessments of Website Value. *Journal of Management Information Systems* 32, 1 (2015), 225–260. https://doi.org/10.1080/07421222.2015.1029394 arXiv:https://doi.org/10.1080/07421222.2015.1029394

[10] Adrian Borrego, Jorge Latorre, Mariano Alcañiz, and Roberto Llorens. 2019. Reliability of the Empatica E4 wristband to measure electrodermal activity to emotional stimuli. In *2019 International Conference on Virtual Rehabilitation (ICVR)*. 1–2. https://doi.org/10.1109/ICVR46560.2019.8994546

[11] Wolfram Boucsein. 2012. *Electrodermal activity*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4614-1126-0

[12] Ming M. Boyer. 2021. Aroused Argumentation: How the News Exacerbates Motivated Reasoning. *The International Journal of Press/Politics* (2021), 19401612211010577. https://doi.org/10.1177/19401612211010577 arXiv:https://doi.org/10.1177/19401612211010577

[13] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[14] Andreas Bulling and Thorsten O. Zander. 2014. Cognition-Aware Computing. *IEEE Pervasive Computing* 13, 3 (2014), 80–83. https://doi.org/10.1109/MPRV.2014.42

[15] Andreas Bulling and Thorsten O. Zander. 2014. Cognition-Aware Computing. *IEEE Pervasive Computing* 13, 3 (2014), 80–83. https://doi.org/10.1109/MPRV.2014.42

[16] Richard B. Buxton, Kâmil Uludağ, David J. Dubowitz, and Thomas T. Liu. 2004. Modeling the hemodynamic response to brain activation. *NeuroImage* 23 (2004), S220–S233. https://doi.org/10.1016/j.neuroimage.2004.07.013 Mathematics in Brain Imaging.

[17] Carole Cadwalladr. 2017. The great British Brexit robbery: how our democracy was hijacked. *The Guardian* 7 (2017).

[18] Filipe R Campante and Daniel A Hojman. 2013. Media and polarization: Evidence from the introduction of broadcast TV in the United States. *Journal of Public Economics* 100 (2013), 79–92.

[19] Pew Reseach Center. 2014. Political polarization in the american public. *Retrieved September* 2 (2014), 2019.

[20] Russ Clay, Jessica M. Barber, and Natalie J. Shook. 2013. Techniques for Measuring Selective Exposure: A Critical Review. *Communication Methods and Measures* 7, 3-4 (2013), 147–171. https://doi.org/10.1080/19312458.2013.813925 arXiv:https://doi.org/10.1080/19312458.2013.813925

[21] Chad Cooper. 2012. The immigration debate in Australia: from federation to World War One. (2012).

[22] Michael E Dawson, Anne M Schell, and Diane L Filion. 2017. The electrodermal system. (2017).

[23] Kristin Demetrious. 2019. 'Energy wars': Global PR and public debate in the 21st century. *Public Relations Inquiry* 8, 1 (2019), 7–22. https://doi.org/10.1177/2046147X18804283 arXiv:https://doi.org/10.1177/2046147X18804283

[24] Tilman Dingler, Benjamin Tag, David A. Eccles, Niels van Berkel, and Vassilis Kostakos. 2022. Method for Appropriating the Brief Implicit Association Test to Elicit Biases in Users. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association

for Computing Machinery, New York, NY, USA, Article 243, 16 pages. https://doi.org/10.1145/3491102.3517570

[25] Tilman Dingler, Benjamin Tag, Evangelos Karapanos, Koichi Kise, and Andreas Dengel. 2020. Workshop on Detection and Design for Cognitive Biases in People and Computing Systems. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3334480.3375159

[26] Tilman Dingler, Benjamin Tag, Philipp Lorenz-Spreen, Andrew W. Vargo, Simon Knight, and Stephan Lewandowsky. 2021. Workshop on Technologies to Support Critical Thinking in an Age of Misinformation. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 105, 5 pages. https://doi.org/10.1145/3411763.3441350

[27] Tilman Dingler, Benjamin Tag, and Andrew Vargo. 2022. Technologies to Support Critical Thinking in an Age of Misinformation (Dagstuhl Seminar 22172). *Dagstuhl Reports* 12, 4 (2022), 72–95. https://doi.org/10.4230/DagRep.12.4.72

[28] Gabriele Donzelli, Giacomo Palomba, Ileana Federigi, Francesco Aquino, Lorenzo Cioni, Marco Verani, Annalaura Carducci, and Pierluigi Lopalco. 2018. Misinformation on vaccination: A quantitative analysis of YouTube videos. *Human Vaccines & Immunotherapeutics* 14, 7 (2018), 1654–1659. https://doi.org/10.1080/21645515.2018.1454572 arXiv:https://doi.org/10.1080/21645515.2018.1454572 PMID: 29553872.

[29] Lauren Feldman, Natalie Jomini Stroud, Bruce Bimber, and Magdalena Wojcieszak. 2013. Assessing Selective Exposure in Experiments: The Implications of Different Methodological Choices. *Communication Methods and Measures* 7, 3-4 (2013), 172–194. https://doi.org/10.1080/19312458.2013.813923 arXiv:https://doi.org/10.1080/19312458.2013.813923

[30] Leon Festinger. 1957. *A theory of cognitive dissonance*. Vol. 2. Stanford university press.

[31] Frank Fishburn, Megan Norr, Andrei Medvedev, and Chandan Vaidya. 2014. Sensitivity of fNIRS to cognitive state and load. *Frontiers in Human Neuroscience* 8 (2014). https://doi.org/10.3389/fnhum.2014.00076

[32] Frank A. Fishburn, Ruth S. Ludlum, Chandan J. Vaidya, and Andrei V. Medvedev. 2019. Temporal Derivative Distribution Repair (TDDR): A motion correction method for fNIRS. *NeuroImage* 184 (2019), 171–179. https://doi.org/10.1016/j.neuroimage.2018.09.025

[33] Ian Flaherty and Jennifer Wilkinson. 2020. Marriage equality in Australia: The 'no' vote and symbolic violence. *Journal of Sociology* 56, 4 (2020), 664–674. https://doi.org/10.1177/1440783320969882 arXiv:https://doi.org/10.1177/1440783320969882

[34] Rudolf Flesch. 1979. How to write plain English. *University of Canterbury. Available at http://www. mang. canterbury. ac. nz/writing_guide/writing/flesch. shtml.[Retrieved 5 February 2016]* (1979).

[35] Society for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures. 2012. Publication recommendations for electrodermal measurements. *Psychophysiology* 49, 8 (2012), 1017–1034. https://doi.org/10.1111/j.1469-8986.2012.01384.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8986.2012.01384.x

[36] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[37] R. Kelly Garrett. 2009. Echo chambers online?: Politically motivated selective exposure among Internet news users. *Journal of Computer-Mediated Communication* 14, 2 (01 2009), 265–285. https://doi.org/10.1111/j.1083-6101.2009.01440.x arXiv:https://academic.oup.com/jcmc/article-pdf/14/2/265/21491614/jjcmcom0265.pdf

[38] Eija Haapalainen, SeungJun Kim, Jodi F. Forlizzi, and Anind K. Dey. 2010. Psycho-Physiological Measures for Assessing Cognitive Load. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing* (Copenhagen, Denmark) *(UbiComp '10)*. Association for Computing Machinery, New York, NY, USA, 301–310. https://doi.org/10.1145/1864349.1864395

[39] Martie G Haselton, Daniel Nettle, and Damian R Murray. 2015. The evolution of cognitive bias. *The handbook of evolutionary psychology* (2015), 1–20.

[40] Dominic Heger, Reinhard Mutter, Christian Herff, Felix Putze, and Tanja Schultz. 2013. Continuous Recognition of Affective States by Functional Near Infrared Spectroscopy Signals. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. 832–837. https://doi.org/10.1109/ACII.2013.156

[41] Vincent F. Hendricks and Mads Vestergaard. 2019. *Reality Lost*. Number 978-3-030-00813-0 in Springer Books. Springer. https://doi.org/10.1007/978-3-030-00813-0

[42] Martin Hilbert. 2012. Toward a synthesis of cognitive biases: how noisy information processing can bias human decision making. *Psychological bulletin* 138, 2 (March 2012), 211—237. https://doi.org/10.1037/a0025940

[43] Leanne Hirshfield, Phil Bobko, Alex Barelka, Natalie Sommer, and Senem Velipasalar. 2019. Toward Interfaces that Help Users Identify Misinformation Online: Using fNIRS to Measure Suspicion. *Augmented Human Research* 4, 1 (Dec. 2019), 1. https://doi.org/10.1007/s41133-019-0011-8

[44] J.M. Inhofe. 2012. *The Greatest Hoax: How the Global Warming Conspiracy Threatens Your Future*. WND Books. https://books.google.co.th/books?id=Ry7SygAACAAJ

[45] Kathleen Hall Jamieson and Joseph N Cappella. 2008. *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press.

[46] Dietmar Jannach and Michael Jugovac. 2019. Measuring the Business Value of Recommender Systems. *ACM Trans. Manage. Inf. Syst.* 10, 4, Article 16 (Dec 2019), 23 pages. https://doi.org/10.1145/3370082

[47] E Jonas, S Schulz-Hardt, D Frey, and N Thelen. 2001. Confirmation bias in sequential information search after preliminary decisions: an expansion of dissonance theoretical research on selective exposure to information. *Journal of personality and social psychology* 80, 4 (April 2001), 557—571. https://doi.org/10.1037//0022-3514.80.4.557

[48] Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.

[49] Jonas T Kaplan, Sarah I Gimbel, and Sam Harris. 2016. Neural correlates of maintaining one's political beliefs in the face of counterevidence. *Scientific reports* 6 (December 2016), 39589. https://doi.org/10.1038/srep39589

[50] Toshinori Kato, Atsushi Kamei, Sachio Takashima, and Takeo Ozaki. 1993. Human Visual Cortical Function during Photic Stimulation Monitoring by Means of near-Infrared Spectroscopy. *Journal of Cerebral Blood Flow & Metabolism* 13, 3 (1993), 516–520. https://doi.org/10.1038/jcbfm.1993.66 arXiv:https://doi.org/10.1038/jcbfm.1993.66 PMID: 8478409.

[51] Joseph T Klapper. 1960. The effects of mass communication. (1960).

[52] Jan Kleinnijenhuis. 2008. *Negativity*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781405186407.wbiecn005 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781405186407.wbiecn005

[53] Silvia Knobloch-Westerwick and Jingbo Meng. 2009. Looking the Other Way: Selective Exposure to Attitude-Consistent and Counterattitudinal Political Information. *Communication Research* 36, 3 (2009), 426–448. https://doi.org/10.1177/0093650209333030 arXiv:https://doi.org/10.1177/0093650209333030

[54] Silvia Knobloch-Westerwick and Jingbo Meng. 2011. Reinforcement of the Political Self Through Selective Exposure to Political Messages. *Journal of Communication* 61, 2 (04 2011), 349–368. https://doi.org/10.1111/j.1460-2466.2011.01543.x arXiv:https://academic.oup.com/joc/article-pdf/61/2/349/22324723/jjnlcom0349.pdf

[55] Silvia Knobloch-Westerwick, Cornelia Mothes, and Nick Polavin. 2020. Confirmation Bias, Ingroup Bias, and Negativity Bias in Selective Exposure to Political Information. *Communication Research* 47, 1 (2020), 104–124. https://doi.org/10.1177/0093650217719596 arXiv:https://doi.org/10.1177/0093650217719596

[56] Asher Koriat, Sarah Lichtenstein, and Baruch Fischhoff. 1980. Reasons for confidence. *Journal of Experimental Psychology: Human learning and memory* 6, 2 (1980), 107.

[57] Arthur E Kramer. 1990. Physiological Metrics of Mental Workload: A Review of Recent Progress.

[58] Katherine Kricorian, Rachel Civen, and Ozlem Equils. 2022. COVID-19 vaccine hesitancy: misinformation and perceptions of vaccine safety. *Human Vaccines & Immunotherapeutics* 18, 1 (2022), 1950504. https://doi.org/10.1080/21645515.2021.1950504 arXiv:https://doi.org/10.1080/21645515.2021.1950504 PMID: 34325612.

[59] Timur Kuran. 1997. *Private truths, public lies: The social consequences of preference falsification*. Harvard University Press.

[60] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[61] Thomas J. Leeper and Rune Slothuus. 2014. Political Parties, Motivated Reasoning, and Public Opinion Formation. *Political Psychology* 35, S1 (2014), 129–156. https://doi.org/10.1111/pops.12164 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/pops.12164

[62] Gilat Levy and Ronny Razin. 2019. Echo Chambers and Their Effects on Economic and Political Outcomes. *Annual Review of Economics* 11, 1 (2019), 303–328. https://doi.org/10.1146/annurev-economics-080218-030343 arXiv:https://doi.org/10.1146/annurev-economics-080218-030343

[63] Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest* 13, 3 (2012), 106–131. https://doi.org/10.1177/1529100612451018 arXiv:https://doi.org/10.1177/1529100612451018 PMID: 26173286.

[64] Stephan Lewandowsky, Gilles E Gignac, and Klaus Oberauer. 2013. The role of conspiracist ideation and worldviews in predicting rejection of science. *PloS one* 8, 10 (2013), e75637.

[65] Scott O. Lilienfeld, Rachel Ammirati, and Kristin Landfield. 2009. Giving Debiasing Away: Can Psychological Research on Correcting Cognitive Errors Promote Human Welfare? *Perspectives on Psychological Science* 4, 4 (2009), 390–398. https://doi.org/10.1111/j.1745-6924.2009.01144.x arXiv:https://doi.org/10.1111/j.1745-6924.2009.01144.x PMID: 26158987.

[66] S. Lloyd-Fox, A. Blasi, and C.E. Elwell. 2010. Illuminating the developing brain: The past, present and future of functional near infrared spectroscopy. *Neuroscience & Biobehavioral Reviews* 34, 3 (2010), 269–284. https://doi.org/10.1016/j.

neubiorev.2009.07.008

[67] Regan L. Mandryk, M. Stella Atkins, and Kori M. Inkpen. 2006. A Continuous and Objective Evaluation of Emotional Experience with Interactive Play Environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montréal, Québec, Canada) *(CHI '06)*. Association for Computing Machinery, New York, NY, USA, 1027–1036. https://doi.org/10.1145/1124772.1124926

[68] Franziska Marquart. 2016. Selective Exposure in the Context of Political Advertising: A Behavioral Approach Using Eye-Tracking Methodology. (2016), 20. https://ijoc.org/index.php/ijoc/article/view/4415

[69] John Martindale, John Mayhew, Jason Berwick, Myles Jones, Chris Martin, Dave Johnston, Peter Redgrave, and Ying Zheng. 2003. The Hemodynamic Impulse Response to a Single Neural Event. *Journal of Cerebral Blood Flow & Metabolism* 23, 5 (2003), 546–555. https://doi.org/10.1097/01.WCB.0000058871.46954.2B arXiv:https://doi.org/10.1097/01.WCB.0000058871.46954.2B PMID: 12771569.

[70] Daina Mazutis and Anna Eckardt. 2017. Sleepwalking into Catastrophe: Cognitive Biases and Corporate Climate Change Inertia. *California Management Review* 59, 3 (2017), 74–108. https://doi.org/10.1177/0008125617707974 arXiv:https://doi.org/10.1177/0008125617707974

[71] Jennifer McCoy, Tahmina Rahman, and Murat Somer. 2018. Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities. *American Behavioral Scientist* 62, 1 (2018), 16–42.

[72] Miriam J. Metzger, Ethan H. Hartsell, and Andrew J. Flanagin. 2020. Cognitive Dissonance or Credibility? A Comparison of Two Theoretical Explanations for Selective Exposure to Partisan News. *Communication Research* 47, 1 (2020), 3–28. https://doi.org/10.1177/0093650215613136 arXiv:https://doi.org/10.1177/0093650215613136

[73] Nir Milstein and Ilanit Gordon. 2020. Validating measures of electrodermal activity and heart rate variability derived from the empatica E4 utilized in research settings that involve interactive dyadic states. *Frontiers in Behavioral Neuroscience* 14 (2020), 148. https://doi.org/10.3389/fnbeh.2020.00148

[74] Randall K. Minas, Robert F. Potter, Alan R. Dennis, Valerie Bartelt, and Soyoung Bae. 2014. Putting on the Thinking Cap: Using NeuroIS to Understand Information Processing Biases in Virtual Teams. *Journal of Management Information Systems* 30, 4 (2014), 49–82. https://doi.org/10.2753/MIS0742-1222300403 arXiv:https://doi.org/10.2753/MIS0742-1222300403

[75] Raymond S. Nickerson. 1998. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology* 2, 2 (1998), 175–220. https://doi.org/10.1037/1089-2680.2.2.175 arXiv:https://doi.org/10.1037/1089-2680.2.2.175

[76] Fateme Nikseresht, Runze Yan, Rachel Lew, Yingzheng Liu, Rose M. Sebastian, and Afsaneh Doryab. 2021. Detection of Racial Bias from Physiological Responses. In *Advances in Usability, User Experience, Wearable and Assistive Technology*, Tareq Z. Ahram and Christianne S. Falcão (Eds.). Vol. 275. Springer International Publishing, Cham, 59–66. https://doi.org/10.1007/978-3-030-80091-8_8 Series Title: Lecture Notes in Networks and Systems.

[77] Alexander Nussbaumer, Katrien Verbert, Eva-Catherine Hillemann, Michael A. Bedek, and Dietrich Albert. 2016. A Framework for Cognitive Bias Detection and Feedback in a Visual Analytics Environment. In *2016 European Intelligence and Security Informatics Conference (EISIC)*. IEEE, Uppsala, Sweden, 148–151. https://doi.org/10.1109/EISIC.2016.038

[78] Markus Ojala and Gemma C Garriga. 2010. Permutation tests for studying classifier performance. *Journal of machine learning research* 11, 6 (2010).

[79] Anneli Olsen. 2012. The Tobii I-VT fixation filter. *Tobii Technology* 21 (2012), 4–19. https://www.tobiipro.com/siteassets/tobii-pro/learn-and-support/analyze/how-do-we-classify-eye-movements/tobii-pro-i-vt-fixation-filter.pdf

[80] Myrto Pantazi, Scott Hale, and Olivier Klein. 2021. Social and Cognitive Aspects of the Vulnerability to Political Misinformation. *Political Psychology* (2021). https://doi.org/10.1111/pops.12797

[81] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you.* Penguin UK.

[82] Gavin W. Ploger, Johnanna Dunaway, Patrick Fournier, and Stuart Soroka. 2021. The psychophysiological correlates of cognitive dissonance. *Politics and the Life Sciences* 40, 2 (2021), 202–212. https://doi.org/10.1017/pls.2021.15

[83] Gregory A Poland, Robert M Jacobson, et al. 2011. The age-old struggle against the antivaccinationists. *N Engl J Med* 364, 2 (2011), 97–9.

[84] Louise Richardson-Self. 2018. Same-Sex Marriage and the "No" Campaign. *Humanities Australia* 9 (2018), 32–39.

[85] Tobias Richter. 2011. *Cognitive Flexibility and Epistemic Validation in Learning from Multiple Texts.* Springer Netherlands, Dordrecht, 125–140. https://doi.org/10.1007/978-94-007-1793-0_7

[86] Tobias Richter. 2015. Validation and Comprehension of Text Information: Two Sides of the Same Coin. *Discourse Processes* 52, 5-6 (2015), 337–355. https://doi.org/10.1080/0163853X.2015.1025665 arXiv:https://doi.org/10.1080/0163853X.2015.1025665

[87] Tobias Richter and Johanna Maier. 2017. Comprehension of Multiple Documents With Conflicting Information: A Two-Step Model of Validation. *Educational Psychologist* 52, 3 (2017), 148–166. https://doi.org/10.1080/00461520.2017.1322968 arXiv:https://doi.org/10.1080/00461520.2017.1322968

[88] Michael Salter. 2016. Men's Rights or Men's Needs? Anti-Feminism in Australian Men's Health Promotion. *Canadian Journal of Women and the Law* 28, 1 (2016), 69–90.

[89] Emily Saltz, Claire R Leibowicz, and Claire Wardle. 2021. *Encounters with Visual Misinformation and Labels Across Platforms: An Interview and Diary Study to Inform Ecosystem Approaches to Misinformation Interventions.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411763.3451807

[90] Josephine B. Schmitt, Christina A. Debbelt, and Frank M. Schneider. 2018. Too much information? Predictors of information overload in the context of online news exposure. *Information, Communication & Society* 21, 8 (2018), 1151–1167. https://doi.org/10.1080/1369118X.2017.1305427 arXiv:https://doi.org/10.1080/1369118X.2017.1305427

[91] Desirée Schmuck, Miriam Tribastone, Jörg Matthes, Franziska Marquart, and Eva Maria Bergel. 2020. Avoiding the Other Side?: An Eye-Tracking Study of Selective Exposure and Selective Avoidance Effects in Response to Political Advertising. *Journal of Media Psychology* 32, 3 (July 2020), 158–164. https://doi.org/10.1027/1864-1105/a000265

[92] Felix Scholkmann, Stefan Kleiser, Andreas Jaakko Metz, Raphael Zimmermann, Juan Mata Pavia, Ursula Wolf, and Martin Wolf. 2014. A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology. *NeuroImage* 85 (2014), 6–27. https://doi.org/10.1016/j.neuroimage.2013.05.004 Celebrating 20 Years of Functional Near Infrared Spectroscopy (fNIRS).

[93] Cornelia Setz, Bert Arnrich, Johannes Schumm, Roberto La Marca, Gerhard Tröster, and Ulrike Ehlert. 2010. Discriminating Stress From Cognitive Load Using a Wearable EDA Device. *IEEE Transactions on Information Technology in Biomedicine* 14, 2 (2010), 410–417. https://doi.org/10.1109/TITB.2009.2036164

[94] Murray Singer. 2013. Validation in Reading Comprehension. *Current Directions in Psychological Science* 22, 5 (2013), 361–366. https://doi.org/10.1177/0963721413495236 arXiv:https://doi.org/10.1177/0963721413495236

[95] Steven M. Smith, Leandre R. Fabrigar, and Meghan E. Norris. 2008. Reflecting on Six Decades of Selective Exposure Research: Progress, Challenges, and Opportunities. *Social and Personality Psychology Compass* 2, 1 (2008), 464–493. https://doi.org/10.1111/j.1751-9004.2007.00060.x arXiv:https://compass.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1751-9004.2007.00060.x

[96] Natalie Jomini Stroud. 2008. Media Use and Political Predispositions: Revisiting the Concept of Selective Exposure. *Political Behavior* 30, 3 (2008), 341–366. http://www.jstor.org/stable/40213321

[97] Natalie Jomini Stroud. 2010. Polarization and Partisan Selective Exposure. *Journal of Communication* 60, 3 (08 2010), 556–576. https://doi.org/10.1111/j.1460-2466.2010.01497.x arXiv:https://academic.oup.com/joc/article-pdf/60/3/556/22324536/jjnlcom0556.pdf

[98] Jill M Swirsky and David Jason Angelone. 2016. Equality, empowerment, and choice: what does feminism mean to contemporary women? *Journal of Gender Studies* 25, 4 (2016), 445–460.

[99] Michael Sülflow, Svenja Schäfer, and Stephan Winter. 2019. Selective attention in the news feed: An eye-tracking study on the perception and selection of political news posts on Facebook. *New Media & Society* 21, 1 (Jan. 2019), 168–190. https://doi.org/10.1177/1461444818791520

[100] Charles S. Taber and Milton Lodge. 2006. Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science* 50, 3 (2006), 755–769. https://doi.org/10.1111/j.1540-5907.2006.00214.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-5907.2006.00214.x

[101] Animals Australia Team. 2019. Surge in Aussies eating vegetarian continues. https://animalsaustralia.org/latest-news/study-shows-surge-in-aussies-eating-veg/

[102] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *science* 185, 4157 (1974), 1124–1131.

[103] Nicholas A. Valentino, Antoine J. Banks, Vincent L. Hutchings, and Anne K. Davis. 2009. Selective Exposure in the Internet Age: The Interaction between Anxiety and Information Utility. *Political Psychology* 30, 4 (2009), 591–613. http://www.jstor.org/stable/25655419

[104] Boris M. Velichkovsky and John Paulin Hansen. 1996. New Technological Windows into Mind: There is More in Eyes and Brains for Human-Computer Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, British Columbia, Canada) *(CHI '96)*. Association for Computing Machinery, New York, NY, USA, 496–503. https://doi.org/10.1145/238386.238619

[105] Micah N. Villarreal, Alexander J. Kamrud, and Brett J. Borghetti. 2019. Confirmation Bias Estimation from Electroencephalography with Machine Learning. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 63, 1 (2019), 73–77. https://doi.org/10.1177/1071181319631208 arXiv:https://doi.org/10.1177/1071181319631208

[106] Nico Voigtländer and Hans-Joachim Voth. 2015. Nazi indoctrination and anti-Semitic beliefs in Germany. *Proceedings of the National Academy of Sciences* 112, 26 (2015), 7931–7936. https://doi.org/10.1073/pnas.1414822112 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1414822112

[107] Michelle W. Voss. 2016. Chapter 9 - The Chronic Exercise–Cognition Interaction: fMRI Research. In *Exercise-Cognition Interaction*, Terry McMorris (Ed.). Academic Press, San Diego, 187–209. https://doi.org/10.1016/B978-0-12-800778-5.00009-8

[108] Kimberlee Weaver, Stephen M Garcia, Norbert Schwarz, and Dale T Miller. 2007. Inferring the popularity of an opinion from its familiarity: a repetitive voice can sound like a chorus. *Journal of personality and social psychology* 92, 5 (2007), 821.

[109] Chris Wells, Katherine J Cramer, Michael W Wagner, German Alvarez, Lewis A Friedland, Dhavan V Shah, Leticia Bode, Stephanie Edgerly, Itay Gabay, and Charles Franklin. 2017. When we stop talking politics: The maintenance and closing of conversation in contentious times. *Journal of Communication* 67, 1 (2017), 131–157.

[110] Drew Westen, Pavel S. Blagov, Keith Harenski, Clint Kilts, and Stephan Hamann. 2006. Neural Bases of Motivated Reasoning: An fMRI Study of Emotional Constraints on Partisan Political Judgment in the 2004 U.S. Presidential Election. *Journal of Cognitive Neuroscience* 18, 11 (11 2006), 1947–1958. https://doi.org/10.1162/jocn.2006.18.11.1947 arXiv:https://direct.mit.edu/jocn/article-pdf/18/11/1947/1935646/jocn.2006.18.11.1947.pdf

[111] Mark P Zanna and Joel Cooper. 1974. Dissonance and the pill: an attribution approach to studying the arousal properties of dissonance. *Journal of personality and social psychology* 29, 5 (1974), 703. https://doi.org/10.1037/h0036651

[112] Arne Zillich and Lars Guenther. 2021. Selective Exposure to Information on the Internet: Measuring Cognitive Dissonance and Selective Exposure with Eye-Tracking. *International Journal of Communication* 15, 0 (2021). https://ijoc.org/index.php/ijoc/article/view/14184

[113] Fabiana Zollo and Walter Quattrociocchi. 2018. *Misinformation Spreading on Facebook*. Springer International Publishing, Cham, 177–196. https://doi.org/10.1007/978-3-319-77332-2_10

4.3 Individual Contributions Towards Article II

I acknowledge that, for the first study, Xiuge Chen recruited participants, conducted the study, and collected the data. I performed a post-hoc analysis of the collected data from the first study, which informed the design of the second study. For the second study, I revised the study design accordingly, recruited participants, collected data, and performed data analysis.

4.4 Chapter Reflection

Cognitive biases happen in human-computer interaction without the user's awareness. Therefore, it is a challenge to detect their occurrences, capture their effects on the interaction, and precisely mitigate the undesired ramifications. In this chapter, addressing **(RQ 2)**, we discuss tools and methods to quantify the occurrences of cognitive biases. We operationalise the scenario of information consumption when individuals likely encounter content expressing opinions on a polarising topic, such as abortion rights or same-sex marriage. Therefore, users may rely on the ideological alignment between their beliefs and the stance of the content as their *heuristics*. As a result, cognitive biases surface when they interact with polarising content.

We set up two user studies (Study 1 and 2) that exposed users to different polarising statements that were either congruent or dissenting from their existing beliefs. During such exposure, we captured behavioural (dwelling time, fixation, and saccade) and physiological (skin conductance level and hemodynamic activity) expressions. Study 1 suggests inconclusive findings because its study design consists of major confounds, such as the continuous presentation of stimulus materials, the influence of cognitive demands on the dwelling time, or image stimuli which are context-specific. Therefore, we revised the study design and deployed Study 2, similar to Study 1, with three significant improvements: we placed an inter-stimuli interval of 15 seconds to allow sufficient delays for physiological expressions; we employed only text stimuli and increased the number of stimulus exposures (from 32 to 64 stimuli); and we collected the users' topic interest, topic familiarity, and three in-study question responses (Q1: participant-stimulus ideological congruence, Q2: likelihood to share the stimulus on one's social media, and Q3: effort spent reading the stimulus) to support the internal validity of the study.

The findings from Study 2 are three-fold. First of all, we found that participants tended to spend more time but reported less reading effort (Q3) reading ideologically dissenting opinions than congruent ones. Secondly (and thirdly), we found significant effects of ideological congruence on hemodynamic activity in individuals who *exhibited lower topic interest*. We observed a non-significant trend in the skin conductance levels. Our results are supported by Richter's two-step model of processing conflicting information [138, 139], suggesting individuals economise mental effort on information that confirms their beliefs while *offloading* their information processing when encountering ideological dissenting information. Richter's model also explains our findings that individuals choose to offload information processing when they have low interest in the topic, thus resulting in significant effects of ideological congruence on the hemodynamic activity.

To conclude this chapter, we show that physiological measurements offer a reliable means to detect the occurrences of cognitive biases in HCI. We cross-validated our findings between self-report measures, user behaviours, and physiological expressions. Our results also point out that user-related factors like topic interest can influence the effects of cognitive biases. It suggests that while cognitive biases are observable, their manifestation is not single-dimensional. In other words, there are a plethora of external factors that influence the effects of cognitive biases. We discuss one major group of such influencing factors – the so-called *bias susceptibility* – in the subsequent chapter.

Chapter 5

Understanding Bias Susceptibility

5.1 Introduction

Although cognitive biases influence how humans process and perceive information, they do not manifest in every individual, context, and scenario. In the previous chapter, we found that the effects of ideological congruency only occurred in individuals who exhibited low topic interest. This finding can be explained by research in psychology. For example, Richter's two-step model of validation [138, 139] suggests that individuals tend to rely on heuristics when processing information; however, if they have relevant background knowledge about the topic, they are more likely to process the information in a more objective manner. Similarly, the elaboration likelihood model of persuasion [20] and the dual-process theory [87] suggest that individuals may shift to analytical thinking when they have enough ability to engage with the information. Implied by our findings and theories in psychology, cognitive biases manifest conditionally depending on the *human* and the *environment* around them. In the context of human-computer interaction, the effects of cognitive biases are subject to the user's *cognitive bias susceptibility*, i.e., factors pertaining to *user* and *system-interaction* context.

Prior research showed that the same factors influencing the effects of cognitive biases also confound the effectiveness of interventions to mitigate cognitive biases. In psychology, Lilienfeld et al. [111] suggested that individual differences in working memory and intelligence could influence an individual's receptibility to debiasing interventions. In the same vein, HCI scholars commented that there is no one-size-fits-all intervention because a user's reaction to interventions is affected by various individual and contextual factors [1, 21, 61, 143]. Aghajari et al. [1] suggested that different individuals have different mental models of interacting with information, and, therefore, factors that drive their cognitive biases are different. Limited research has explored how factors related to the user and interaction context influence the effects of cognitive biases. Especially, our understanding of what *amplifies* or *hinders* the effects of cognitive biases in human-computer interaction is insufficient. In other words, it is unclear what factors influence the users' susceptibility to exhibit cognitive biases when interacting with computing systems.

In this chapter, we present a user study investigating the influence of individual and contextual factors on the effects of confirmation bias when individuals seek, interpret, and recall information on a news feed. Primarily, we operationalise confirmation bias as it is one of the most prominent forms of cognitive biases [43, 128]. Confirmation bias refers to the tendency to seek, interpret, and recall predominantly information that confirms one's beliefs [124, 192]. We assessed three relevant scenarios of information consumption where confirmation bias manifests: information seeking, interpretation, and recall [183]. In this study, we exposed participants to a news feed of tweets containing opinions on a polarising topic. The setting is similar to the studies in the previous chapter: each tweet expressed an opinion that either supported or opposed one

ideology. Therefore, we assumed that the tweets would trigger individuals to rely on the alignment between the tweet's ideological stance and their existing beliefs as a heuristic, subsequently triggering confirmation bias. We measured the effects of confirmation bias on the intention for information-seeking, the recall ability, and the subjective perception of the information through three information-consumption tasks: ranking the headline titles according to their preferences, recalling the details of the tweets presented in a news feed, and rating their perception of each individual tweet. To assess the influence of factors on individuals' susceptibility to exhibit confirmation bias, we gauged the predictors from individual measures (the tendency for effortful thinking, need for cognition, bullshit receptibility, and the tendency for political conservative beliefs) and contextual measures (topic interest and the perceived issue strength of each tweet).

The results showed that the tendency for effortful thinking, strong political liberal beliefs, and strong perceived issue strength of the content amplified the effects of confirmation bias. Moreover, the modality of the task also influenced the effects of confirmation bias, as the influence of bias susceptibility predictors varied across different tasks. For example, while we found that the tweet's perceived issue strength and the individual's effortful thinking tendency influenced confirmation bias in information-seeking intention, these factors did not influence confirmation bias in information recall. In sum, we provide empirical contributions to HCI: we suggest that *individual* and *interaction* contexts can both influence how cognitive biases manifest when users interact with computing systems and, therefore, the user's bias susceptibility.

Informed by our findings, we discuss practical and ethical implications for designers of social media platforms to consider the user's bias susceptibility when designing and employing interventions to mitigate undesired effects of cognitive biases. We incorporate the notion of *context-awareness* into intervention designs. Interventions can shift away from one-size-fits-all approaches to personalised, context-aware interventions that consider individual and contextual differences in the interaction. For example, platforms could employ linguistic models to detect the strong stance and sentiment of the content and adjust it towards a more nuanced perspective to help safeguard users from falling victim to confirmation bias. Moreover, platforms could consider individual differences in thinking styles and political belief tendencies, identify users' bias susceptibility, and employ prevention interventions, like psychological inoculation or media literacy education, to train and fortify them against manipulation. Nonetheless, we call for intervention designers to consider ethical consequences arising from the use of bias susceptibility to inform interventions, as it can be viewed as a form of benevolent paternalism (i.e., limiting the user's agency in the best interest of the people). The same bias susceptibility factors can be abused as social engineering applications to manipulate individuals' decision-making without their awareness.

We describe the design and analysis of this study in more detail and discuss its implications in the attached publication, Article III.

5.2 Article III

Nattapat Boonprakong, Saumya Pareek, Benjamin Tag, Jorge Goncalves, and Tilman Dingler. 2025. Assessing Susceptibility Factors of Confirmation Bias in News Feed Reading. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3706598.3713873

Ethics Application ID: 1956072.1, the University of Melbourne Human Research Ethics Committee.

Assessing Susceptibility Factors of Confirmation Bias in News Feed Reading

Nattapat Boonprakong
School of Computing and Information Systems
University of Melbourne
Parkville, Victoria, Australia
nboonprakong@student.unimelb.edu.au

Saumya Pareek
School of Computing and Information Systems
University of Melbourne
Melbourne, Victoria, Australia
spareek@student.unimelb.edu.au

Benjamin Tag
School of Computer Science and Engineering
University of New South Wales
Sydney, New South Wales, Australia
benjamin.tag@unsw.edu.au

Jorge Goncalves
School of Computing and Information Systems
University of Melbourne
Melbourne, Australia
jorge.goncalves@unimelb.edu.au

Tilman Dingler
Industrial Design Engineering
Delft University of Technology
Delft, Netherlands
t.dingler@tudelft.nl

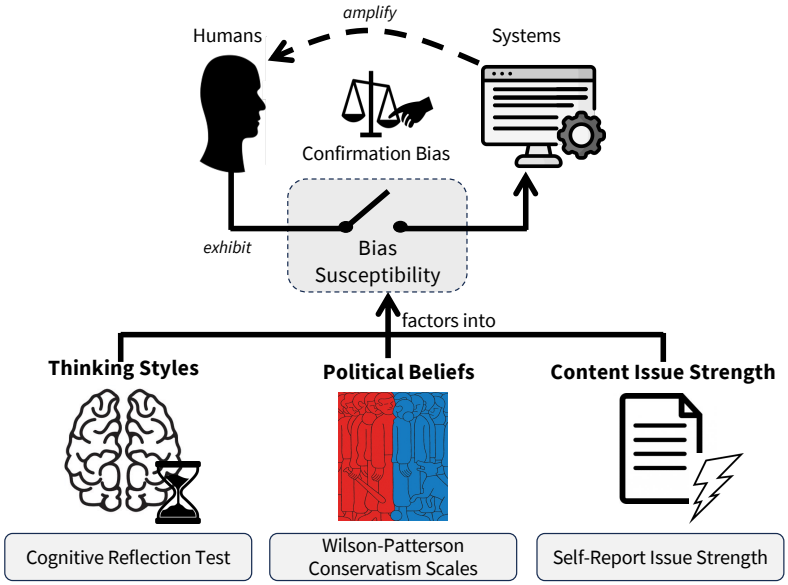


Figure 1: We investigated the human factors influencing susceptibility to confirmation bias in news feed reading and identified the individual's thinking styles, the strength of political beliefs, and the content's perceived issue strength as key contributors.

Abstract

Individuals tend to apply preferences and beliefs as heuristics to effectively sift through the sheer amount of information available online. Such tendencies, however, often result in cognitive biases, which can skew judgment and open doors for manipulation. In this work, we investigate how individual and contextual factors lead to instances of confirmation bias when seeking, evaluating, and recalling polarising information. We conducted a lab study, in which we exposed participants to opinions on controversial issues through a Twitter-like news feed. We found that low-effortful thinking, strong political beliefs, and content conveying a strong issue amplify the occurrences of confirmation bias, leading to skewed information processing and recall. We discuss how the adverse effects of confirmation bias can be mitigated by taking bias-susceptibility into account. Specifically, social media platforms could aim to reduce strong expressions and integrate media literacy-building mechanisms, as low-effortful thinking styles and strong political beliefs render individuals especially susceptible to cognitive biases.

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713873

Keywords

cognitive bias, confirmation bias, individual difference, recall, social media

ACM Reference Format:

Nattapat Boonprakong, Saumya Pareek, Benjamin Tag, Jorge Goncalves, and Tilman Dingler. 2025. Assessing Susceptibility Factors of Confirmation Bias in News Feed Reading. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3706598.3713873

1 Introduction

Given the sheer amount of information available online, individuals apply **cognitive biases** as their "rule of thumb" to effectively skim through this information [5]. However, cognitive biases often skew our judgment and prompt us to give up analytical thinking [105]. These biases, therefore, can be harmful as they open doors for manipulation. Misinformation and conspiracy theories, for example, tend to trigger our cognitive biases to create more engagement about controversial, divisive issues that touch on people's preexisting beliefs (*e.g.*, politics and human rights) [18, 35]. Social engineering attacks also try to tap into individuals' cognitive biases to steer their behaviours [14]. In the Cambridge Analytica scandal [10], for example, people's personal tendencies on social media were (mis-)used to target their cognitive vulnerabilities and subsequently sway their opinion-making all without their awareness.

When consuming online information, individuals often rely on their existing preferences and beliefs as cognitive strategies – shaped by their previous experience of the world [56, 141] – to effectively process information presented to them. However, this often results in cognitive biases, which prompt individuals to see only what they want to see without carefully inspecting the content piece [6, 117]: for instance, *confirmation bias* increases people's tendency to predominantly seek, interpret, and recall information that aligns with their beliefs [96]; and *cognitive dissonance* leads to the avoidance of information deemed incongruent to one's beliefs [42]. These biases and tendencies can be triggered by the information content that conveys an ideologically polarising issue and, more importantly, can be amplified by algorithmic information curation, which tends to optimise and cater predominantly to the users' preferences and beliefs, even when these may be misinformed [7, 59, 80].

A variety of approaches has been proposed to mitigate cognitive biases using behavioural interventions, such as *nudging* [17] or *boosting* [93]. However, research has suggested that they are not always effective [12, 114]. Specifically, there is no one-size-fits-all solution for debiasing because a person's reaction to debiasing approaches is affected by a variety of individual and contextual factors [2, 49, 89, 110]. Moreover, research has highlighted **individual factors** that dictate why some individuals might be particularly susceptible to exhibiting cognitive biases while others are more resistant. Research in psychology has suggested that individual differences influence how people perform reasoning [127] and how receptive they are to cognitive bias interventions [40, 89]. In the realm of misinformation, studies [33, 104, 134] have shown that individual differences in effortful thinking styles, assessed by the Cognitive Reflection Test (CRT) [44], can predict an individual's susceptibility to misinformation. At the same time, people may react to information stimuli differently, depending on the context and situation of the interaction. Furthermore, there are **contextual factors**, which describe the relationship between the user and the triggers of cognitive biases. For example, an individual's interest and involvement in the topic have been shown to influence how they interact with information [86, 87] and, more importantly, the extent to which they are susceptible to cognitive biases [11, 146]. Moreover, several studies have suggested that attitude strength and attention are essential in activating cognitive biases [1, 109].

By studying susceptibility factors for cognitive biases, we can pave the way for designing more effective bias mitigation techniques that adapt to individuals and interaction contexts. Yet, limited studies have investigated how these factors come into play in human-computer interaction. In this research, we tackle the question – "*How do individual and contextual factors influence the occurrences of cognitive biases* ?" We assess the interplay of individual and contextual factors that influence the manifestation of cognitive biases and the degree to which how people are susceptible to them when interacting with computing systems. To closer study this, we operationalise *confirmation bias*, which presents a tendency of people to rely on their attitudes and ideological beliefs when seeking, interpreting, and recalling information [96, 99, 144]. Confirmation bias is one of the most prominent forms of cognitive bias and is highly prevalent in information consumption [5, 61, 146]. We conducted a user study that exposed participants to information on controversial, divisive issues. We asked them to rank headlines according to reading preference, read and recall a news feed, and evaluate the reliability of individual tweet-like information. Through regression analyses, we examined the interaction effects between confirmation bias, *i.e.*, reliance on prior beliefs when evaluating information, and bias susceptibility factors. Specifically, we investigated which and how individual and contextual factors might *amplify* the effects of confirmation bias in three information consumption scenarios: information-seeking intention, information recall, and information interpretation.

We found that the tendency for effortful thinking, strong political beliefs, and strong issue strength of the content (perceived by the study participants) amplified the effects of confirmation bias (Figure 1). Specifically, ideologically polarised information tended to stand out more in people's memories, especially when it confirms their ideological beliefs. Individuals holding strong political beliefs tended to let their information consumption behaviours be guided by their attitudes. In addition, we found that the design and modality of the task influenced the occurrence of confirmation bias. In summary, this work makes the following contributions:

(1) We present an empirical investigation into individual and contextual factors that make users susceptible to falling for their confirmation bias during information consumption.
(2) We provide a discussion of how interventions can be designed to effectively mitigate the effects of confirmation bias by taking into account the bias susceptibility factors based on the characteristics of the users, the content, and the interaction between them. We also discuss ethical and practical implications for media platforms when incorporating bias susceptibility into intervention designs.

2 Related Work

Our work is grounded in research on behavioural psychology in the context of recent discussions in Human-Computer Interaction (HCI) regarding the interplay of cognitive biases, computing systems, and their users.

2.1 Cognitive Bias

In the 1950s, psychologist Herbert Simon proposed the concept of *bounded rationality* – human rationality is inherently limited [121]. Given the complexity of the world and information present to them, humans apply *mental shortcuts* or *heuristics* to make faster but less deliberate decisions. Amos Tversky and Daniel Kahneman later extended Simon's concept of bounded rationality into the notion of *cognitive bias* [141], where they laid out how mental shortcuts systematically skew humans behaviours from the norm of rational judgment without their awareness. Psychologists and behavioural scientists have documented different forms of cognitive biases. For example, anchoring bias presents a tendency where people rely on the first piece of information they see [141], or availability bias makes individuals rely on information that is mostly available to them [140].

Subsequent research in psychology has augmented the original definition of cognitive biases as *features of the human mind* to cope with the complexity of the world. The prominent psychologist Gerd Gigerenzer viewed that humans apply heuristics as cognitive strategies to effectively make fast decisions [48]. Lieder et al. [88] argued that cognitive biases are mechanisms that humans use to make optimal decisions under their limited cognitive resources. From the lens of evolution psychology, Haselton et al. [55, 56] suggested cognitive biases are inherent mechanisms humans employ as part of their survival and natural adaptation. More importantly, individuals form heuristics or their "rules of thumb" based on the beliefs and preferences they learned from past experiences of the world.

In the realm of online information consumption, users generally have cognitive biases as their inherent, unconscious cognitive strategies for effectively skimming through the sheer volume of information on their news feed and stopping at the news piece of their interest. Different forms of cognitive biases, therefore, affect how humans perceive and evaluate information. For example, confirmation bias [96] and cognitive dissonance [42] prompt individuals to favour information that aligns with their beliefs and avoid what is deemed incompatible. Others, like the continued influence effect [83], make individuals stick to false information although it has been retracted, while negativity bias [75] triggers stronger attentional and emotional responses to information with a negative affect. These cognitive biases become problematic as they enhance the ability of misinformation to deceive people, be disseminated, and persist in memory [6, 18, 117].

Recent discussions in HCI [12, 91] have brought attention to the role of algorithms and recommendation systems in amplifying cognitive biases in users. Different forms of cognitive bias prompt users to seek and expose themselves to information favouring their beliefs. At the same time, recommendation systems optimise on and cater predominantly to the users' preferences and beliefs [9], resulting in amplifying their existing cognitive biases [7, 59, 80]. Without proper intervention, cognitive biases and recommender algorithms together form a self-reinforcing loop, hinder users' ability to make an informed decision, and make them vulnerable to manipulation [3].

A growing body of work has explored how the adverse effects of cognitive biases could be mitigated [12, 110]. Prior research has investigated various debiasing techniques, such as nudging [17, 118, 137] (tapping into people's cognitive biases to shift them towards a desirable behaviour outcome), boosting [93, 111] (nurturing people's metacognitive skills), or decision-support systems [147] (guiding users to make informed, optimal decisions). However, effectively mitigating cognitive biases is challenging, mainly because some individuals are more susceptible to cognitive biases than others [47, 91, 110]. In other words, no one-size-fits-all solution exists: different individuals possess different mental models of interacting with information [2, 49], and, therefore, factors that drive their cognitive biases could be different. Limited research has explored how user- and context-related characteristics come into play with regard to cognitive biases in the context of social media. We review these factors in the following sections, focusing on confirmation bias.

2.2 Confirmation Bias

Confirmation bias presents a tendency to seek, interpret, and recall evidence in a way that they are partial to beliefs, preferences, or hypothesis in hand [96, 99, 148]. It is a long-established phenomenon in psychology [74] and one of the most prominent forms of cognitive biases [38, 98]. In his seminal work, Nickerson [96] demonstrated that confirmation bias occurs largely in humans' everyday decision-making without their awareness, for example, the tendency for people to make a hypothesis about number patterns (*i.e.*, number mysticism [148]), the tendency for doctors to find evidence to support their medical diagnosis, or the tendency for jurors to interpret ambiguous evidence pieces in favour of their pre-existing beliefs.

Notably, confirmation bias overlaps with related phenomena in psychology like motivated reasoning [73, 133]. While both emerge from individuals' reliance on the ideological congruence between their beliefs and the information, each pursues a different scope. Confirmation bias is primarily an unconscious cognitive mechanism that reinforces one's existing beliefs. Meanwhile, motivated reasoning refers to a goal-driven tendency (*e.g.*, to defend one's ideology or values [30]) to favour evidence that confirms one's beliefs while rejecting information deemed unfit. The latter is broader in scope as motivated reasoning can also be deliberately driven by goals and emotions in the reasoning process, as well as subconsciously influenced by confirmation bias [54, 68]. While these phenomena pursue different mechanisms, they both influence how individuals seek, perceive, and recall information. In the realm of information consumption, confirmation bias and motivated reasoning have been attributed to political polarisation [68, 133] and the spread of misinformation [21], as well as giving rise to *selective exposure* [46] (known as individuals' tendency to expose themselves to predominantly information that confirms their beliefs [43]).

In this paper, we investigate factors that influence how people are susceptible to cognitive biases through the lenses of confirmation bias. In conjunction, we operationalise (1) confirmation bias

in the context of information consumption as when users *rely* on the congruence between their beliefs and the ideological stance of the content present to them on a news feed, and (2) user- and interaction-context-related characteristics that influence how cognitive biases manifest.

2.3 The Occurrences of Confirmation Bias in Information Consumption

Confirmation bias exists in many stages of information consumption: it prompts individuals to rely on their attitudes and beliefs, affecting how they seek, perceive, and remember information they encounter. The effects of confirmation bias, therefore, distort individuals' psychological expression (*e.g.*, perception [144], cognitive load [119], and recall [45]), behavioural expression (*e.g.*, clicks [134] and attention [131]), or physiological expression (*e.g.*, peripheral and brain signals [11]). Following Vedejová and Čavojová [144], in this work, we operationalise three scenarios where confirmation bias can manifest: information-seeking intention, interpretation, and memory recall. Next, we briefly review related works that have investigated the effects of confirmation bias in each of the information scenarios.

2.3.1 Information Seeking. Research has studied confirmation bias in information seeking from the lenses of selective exposure [20, 43, 74]. Recent research in HCI has studied the same phenomenon in the online context as users tend to exhibit information selection behaviours in favour of belief-congruent information [86, 106, 112, 134]. For example, Liao and Fu [86] and Pothirattanachaikul et al. [106] showed that people clicked to read more predominantly content items that confirmed their beliefs. On the other hand, Tanaka et al. [134] found that users tended to avoid clicking on fact-checking messages that contradict their pre-existing beliefs.

2.3.2 Information Interpretation. Confirmation bias prompts individuals to evaluate congruent information differently from dissenting information. In psychology, Kobayashi [76] found that individuals tended to put more scrutiny on information against their beliefs. Research in HCI has also studied how users are influenced by their beliefs as a heuristic when evaluating information. van Strien et al. [143] conducted an eye-tracking study and found that individuals' strong attitudes can skew how they evaluate the creditability of information on the web. Allen et al. [4] showed that Twitter Birdwatch users preferably challenge fact-checking content from those with whom they disagree politically. In another example, Wischnewski et al. [149] found that individuals tended to perceive Twitter profiles deemed incongruent to their beliefs as bot accounts.

2.3.3 Information Recall. Based on the schema theory, *schemas*, known as the knowledge structure, is built over time from experiences and memories. A memory schema causes different pieces of information to be remembered differently, thus, resulting in the application of confirmation bias [31, 126]. A seminal study [92] reporting on a car crash experiment suggested that the memory of an event can be distorted by the *perception* of the details during the actual event. A small number of studies, however, have investigated how confirmation bias affects the recall of information and produced mixed results. Some studies have suggested that individuals better recall information that supports their attitudes or beliefs [45, 50, 66, 95]. For example, Frost et al. [45] conducted a study where participants were asked to recognise social media posts. They found that the recognition memory for information congruent with their viewpoints was better than that for dissenting information. Meanwhile, some works have found opposite results. For example, in their first study, Lescarret et al. [81] reported that middle school students tended to recall better attitude-inconsistent information, while there was no such effect in university students. Other research suggested no difference in the recall ability for information supporting or opposing one's prior attitudes [62, 130, 144].

2.4 Influencing Factors for the Occurrence of Cognitive Biases

2.4.1 Individual Factors. Research has pointed out several factors governing individuals' tendency to fall victim to cognitive biases. One of the most prominent indicators is the individual difference in *effortful thinking styles*. Cognitive biases are byproducts of using our intuitive, fast System 1 thinking instead of the deliberate but slower System 2 thinking [39, 69]. Research has shown that individual differences in effortful thinking styles correlate with the occurrences of cognitive biases [128, 136, 145] and the discernment of misinformation [8, 85, 104]. Some of the effortful thinking indicators include the Cognitive Reflection Test (CRT) [44], which measures one's tendencies to use intuitive thinking over deliberative thinking, the Need for Cognition Scale (NFC) [15], which gauges the tendency to engage in effortful cognitive activities, and the Bullshit Receptivity Scale (BRS) [102], which reflects the ability to detect *bullshit* or statements with profound meaning. Recent research in HCI has increasingly used these techniques to investigate the relationship between such factors and how people interact with information online [70, 134].

Moreover, political attitudes also affect individuals' receptivity to cognitive bias interventions. Research in psychology has shown that individual differences in political ideology influence how they process information [36, 120]. In addition, studies have shown that people who leaned towards conservative beliefs were more likely to exhibit less reflective thinking [25] and be more resistant to misinformation correction than individuals on the liberal end of the political spectrum [34, 35, 51]. Yet, recent research has argued that this tendency did not hold exclusively for political conservatives, as inclinations for liberal beliefs [37] or any of the political extremes [142] render people susceptible to conspiracy thinking.

2.4.2 Contextual Factors. The occurrences of cognitive biases also depend on the contextual relationship between the user and the information. In one direction, an individual's interest or involvement with the information describes such relationships. This has been highlighted as a moderating or amplifying factor for cognitive biases in prior HCI studies [11, 86, 87, 146]. In behavioural psychology, Richter et al. [108, 109] proposed the Two-Step Model of Validation, which explains that individuals with relevant background knowledge tended to process conflicting information more elaboratively. In other words, they may bypass the use of cognitive biases towards a more balanced way of information processing.

The information's ability to amplify or trigger cognitive biases explains the other end of this contextual relationship as well. Research has pointed out its ability – *e.g.*, strong language and slant – to trigger people's negative emotional responses [94] and attention [1, 115], thus, making them susceptible to believing and sharing false information. Recent works in HCI have also highlighted linguistic and sentimental features that prompt strong emotional responses and potentially trigger people's mental shortcuts [123, 151].

2.5 Our Contributions

People adopt cognitive biases to effectively sift through a large amount of information presented to them when they roam through online platforms. This comes at a cost; cognitive biases prompt their irrationality and make them vulnerable to manipulation. While it is important to mitigate the adverse effects of cognitive biases, recent research in HCI has suggested that bias mitigation is challenging because cognitive biases manifest differently according to various innate user characteristics and contexts of the interaction between users, information, and systems. Our work contributes to HCI research as we investigate the influence of individual and contextual factors on cognitive bias susceptibility. We consider three relevant scenarios of information consumption where confirmation bias manifests: (1) information-seeking intention, (2) interpretation, and (3) recall. Accordingly, we employ three representative tasks for information consumption: (1) headline ranking, (2) individual tweet evaluation, and (3) news feed-free recall. Based on our findings, we shed light on implications for designers and media platforms tailoring effective interventions that consider bias susceptibility in users, information, and their interaction.

Additionally, this work contributes new knowledge to the literature by looking at the influence of bias susceptibility factors on recall ability, on which there has been limited research. In addition, to the best of our knowledge, our work is the first to assess the effects of confirmation bias on memory recall using a *delayed free recall* task, where participants are presented with a sequence of tweet-like information and subsequently, after some delays, asked to recall them in any order.

3 Methodology

We conducted a user study to explore indicators for susceptibility to confirmation bias. Therefore, we exposed participants to a number of tweets stating opinions on controversial topics. Subsequently, they engaged in three tasks: ranking headlines, recalling tweets on a news feed, and evaluating tweets.

3.1 Study Stimuli

During the experiment, we showed tweets containing only textual information (*i.e.*, no images) on three controversial, polarising topics. Each tweet aligned with one end of the ideological spectrum, *i.e.*, supporting (pro) or opposing (con). We picked tweets concerning the following topics: *abortion rights*, *same-sex marriage*, and *vegetarianism*. All chosen topics have been widely debated globally with increasingly polarised viewpoints [23, 57, 101] and lend themselves, therefore, well to our study.

We sourced all tweets from the ProCon.org website[1], which provides facts, opinions, and arguments on various controversial topics on both ends of the ideological spectrum. For example, on the abortion rights issue, the *pro* stance endorses the idea that *abortion should be legal*. In contrast, the *con* stance supports the idea that *abortion should be prohibited*. Table 2 shows pro-con ideology pairs for each topic deployed in this study. In addition, we made sure all tweets were in English and between 40 to 70 words in length to resemble the standard 250-character tweets. Following the approach in related works [11, 29, 112], we hypothesise that the tweets would trigger participants' confirmation biases by making them rely on their pre-existing beliefs when assessing the information.

We gathered eight tweets on each of the three topics, consisting of four pro and four con tweets. Each tweet was presented on the screen with the same font, compact line spacing, alignment, column width, colour text (black), and white background (see Figure 2, middle item). We deployed our stimuli and questionnaire on Qualtrics[2]. We did not provide source information in our stimuli to separate confounds such as source bias [135].

3.2 Study Design

3.2.1 Experimental Design. To study the effects of contextual and individual factors on the occurrences of confirmation bias, we conducted a study with a within-subject design. We measured the effects of confirmation bias (in three scenarios: information-seeking intention, information recall ability, and information interpretation ratings) and investigated the influence of the following predictors: the ideological congruence score between the user and tweet, individual factors (as measured by the Cognitive Reflection Test (CRT), Need For Cognition scores (NFC), Bullshit Receptivity Scales (BRS), and Wilson-Patterson Conservatism Scales (WPCS)), and contextual factors (topic interest, and tweet perceived issue strength). Table 1 summarises all predictor and measurement variables we examined in this study. We determined the required sample size ($N = 42$) by a priori power analysis using *G*Power* [41] with a medium-to-large effect size $f^2 = 0.25$, power $1 - \beta = 0.80$, type I error probability $\alpha = 0.05$, and two predictors.

3.2.2 Participants. We invited 42 participants (16 men, 25 women, and one non-binary) through the university network to join the study in our usability lab. All participants were native or fluent speakers of the English language, and their mean age was 28.51 years ($SD = 8.52$), with the minimum and maximum ages being 19 and 54 years old, respectively. Of 42 participants, 11 held a postgraduate degree, 19 held a bachelor's degree, and the remaining 12 participants had at least 12 years of education.

3.2.3 Procedure. The study took place in a quiet room in our institution's usability lab. We first informed each participant about the objective and procedure of the study and collected their written consent. We seated participants before a screen and asked them to adjust their seating to a comfortable position. Participants responded to a pre-study survey collecting information on individual

[1] www.procon.org
[2] www.qualtrics.com

Table 1: List of the examined predictor and measurement variables.

Variable	Measure	Scale
Predictor Variables	**Ideological Congruence**	
	- Implicit Congruence (Cong_Imp) [77, 86]	Ordinal (-7 to +7)
	Individual Factors	
	- Cognitive Reflection Test score (CRT) [44]	Count of correct responses (0 to 7)
	- Need For Cognition scale (NFC) [15]	Ordinal (5-Likert scale: 1 to 5)
	- Bullshit Receptivity Scale (BRS) [102]	Ordinal (5-Likert scale: 1 to 5)
	- Wilson-Patterson Conservatism Scale (WPCS) [58]	Count of conservative items (−27 to +27)
	Contextual Factors	
	- Topic Interest (Interest) [63, 86]	Ordinal (7-Likert scale: 1 to 7)
	- Stimulus Perceived Issue Strength (Strength)	Ordinal (5-Likert scale: 0.5 to 4.5)
Measurement Variables	**Information-Seeking Intention**	
	- Headline Rank Position	Ordinal (1 to 8)
	Information Interpretation	
	- Information Interpretation scale [73, 144]	Ordinal (5-Likert scale: 1 to 5)
	Information Recall	
	- Recall Ability Score	Ordinal (4-Likert scale: 0 to 3)

Table 2: Topics presented and their ideological ends.

Topic	*Pro* stance	*Con* stance
Vegetarianism	People should become Vegetarian	People should not become Vegetarian
Abortion Rights	Abortion should be legal	Abortion should be prohibited
Same-sex Marriage	Same-sex marriage should be legal	Same-sex marriage should be prohibited

and contextual factors with their demographic information (see Section 3.3.1), after which they were asked to complete the following tasks:

(1) **Ranking Headlines**: Participants were presented with a list of tweet headlines (eight per topic), each of which was a one-sentence snippet of a tweet. The headlines were initially presented in a random order. The participants had to reorder the headlines from what they wanted to read the most (top, first place) to what they wanted to read the least (bottom, last place).
(2) **Reading and Recalling Tweets**: Subsequently, participants were presented with a news feed of eight tweets in a random order. They could scroll up and down to read each tweet. Once they finished reading the tweets, participants were asked to engage in an interruption task where they had to solve seven summations (adding two single-digit numbers). Inspired by previous studies that employed recall tasks [62, 125], the interruption task was introduced to separate participants mentally from the tweets and to reset their working memory, as well as to serve as attention checks. After that, participants responded to a free-recall task, namely "*please write down every viewpoint, aspect, and detail that you can remember from the tweets you have just read.*" Participants were given a maximum of five minutes to type in their responses using a keyboard. They were told to ignore spelling and grammatical errors.
(3) **Evaluating Individual Tweets**: Lastly, participants were again presented with the earlier shown tweets and asked to rate the tweet according to the information interpretation scales and how polarised it was (See Section 3.3.2). Each tweet was presented with the in-study survey on the same page. Once they finished the survey, participants could click "Next" and proceed to the subsequent tweet.

We repeated all three tasks for each topic (abortion rights, same-sex marriage, and vegetarianism) and incorporated eight tweets (four pros and four cons) per topic. The presentation order of topics was randomised, while the tasks were repeated in a fixed order. A short break between each topic allowed participants to relax for at least 15 seconds before proceeding to the next topic. Figure 2 visualises the procedure of our study.

Upon completion, we compensated each participant with a $20 electronic voucher for their time. The study took about an hour to complete and was approved by the Human Research Ethics Committee of the University of Melbourne.

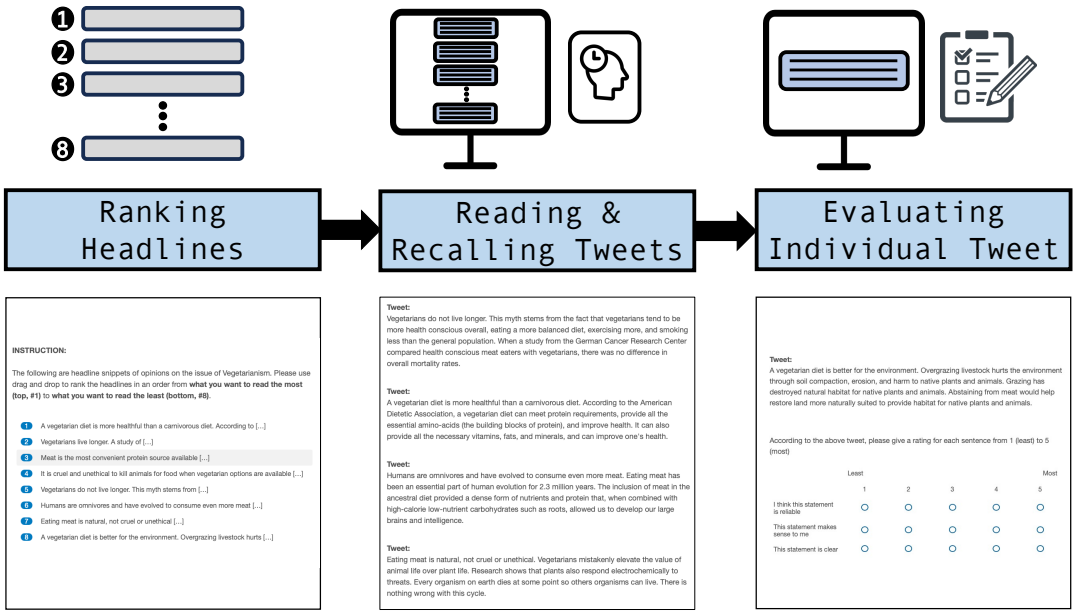


Figure 2: Summary of the study procedure. For each topic, participants completed three tasks in the following order: Headline Ranking, Reading and Recalling Tweets, and Individual Tweet Evaluation. Example screenshots of the tasks are shown below.

3.3 Surveys

3.3.1 Pre-study Survey. We gauged potential individual and contextual factors for cognitive biases through the pre-study survey, which was administered after the participant provided informed consent. The survey started with four tests: **Cognitive Reflection Test (CRT)**, **Need For Cognition scores (NFC)**, **Bullshit Receptivity Scales (BRS)**, and **Wilson-Patterson Conservatism Scales (WPCS)**. We derived the 7-item version of CRT [122] consisting of seven binary-choice questions gauging the participant's tendency to use System 1 over System 2 thinking. For each question, there was one correct answer and one wrong answer. NFC consists of questions asking participants to self-report their tendency to use System 1 thinking on a 7-Likert Scale (1: extremely uncharacteristic to 7: extremely characteristic) [15]. For the BRS, we followed the approach in Pennycook et al. [102], presenting participants with a list of 10 profound statements in English, asking them to rate how profound they thought each statement was on a 5-Likert scale (1: least profound and 5: most profound). Finally, for the WPCS, we presented participants with 27 socio-economic policies and asked them to indicate whether they agreed, disagreed, or felt neutral with each policy. We opted for the scales by Henningham [58], developed exclusively for our study population in Australia. The number of liberal and conservative policies were counterbalanced, and their presentation order was randomised. For each of the four tests, the presentation order was randomised.

In addition, for each of the three topics, we asked participants to rate how they felt about the following statements on a 7-Likert scale (1: strongly disagree to 7: strongly agree): "*this issue is related to my core values*," "*it is important to defend my point of view on this issue*," "*I am interested in learning about this issue*," and "*I desire to know the facts about this issue*." Following the approach in [63, 86], we took the average of their four responses to measure the participants' **topic interest**. We then measured participants' ideological stance on the topic in a self-report question and a word association test. In the self-report question, we asked participants to rate their stance on a continuous scale from 0 to 100 (0: strongly opposing to 100: strongly supporting). In the word association test, participants were presented with each topic name (*i.e.*, abortion rights, same-sex marriage, and vegetarianism) and asked to associate each of them with five pairs of bipolar adjectives: *unfavourable-favourable*, *bad-good*, *unnecessary-necessary*, *harmful-beneficial*, and *unhealthy-healthy*. Following the same test deployed in prior studies [77, 86], we asked participants to choose their stance on a 7-Likert scale (1: negative adjective to 7: positive adjective).

3.3.2 In-study Survey. For each tweet presented, we gauged how participants interpreted the tweet using the information interpretation scales derived from Klaczynski [73]. We asked them to rate the following statements on a 5-Likert scale (1: strongly disagree to 5: strongly agree): "*This tweet is reliable to me*", "*This tweet is clear to me*", and "*This tweet makes sense to me*". In addition, we asked participants to rate how polarised the tweet was according to its expressed ideology. On a 10-Likert scale, we asked them to rate whether the tweet supports the pro or con stances (1: strongly supporting the con stance to 10: strongly supporting the pro stance).

3.4 Predictor Variables

3.4.1 Individual Factors. We derived the following individual factors: CRT, NFC, BRS, and WPCS. We calculated the CRT scores by counting the number of correct responses. Higher CRT, therefore, indicates a higher tendency for effortful thinking [44]. For NFC and BRS, we averaged the internal question items, 6 for NFC and 10

Table 3: Descriptive statistics (Mean, S.D., max, min) of the topic-wise ideological stances and contextual factors collected in our study

Factors	*Abortion Rights*	*Vegetarianism*	*Same-sex Marriage*
Stance_Imp (7-Likert scale: [1..7])	M = 5.13, S.D. = 1.65, max = 7, min = 1	M = 3.42, S.D. = 1.64, max = 6.25, min = 1	M = 4.40, S.D. = 1.96, max = 7, min = 1
Stance_Exp (continuous: [0..100])	M = 86.76, S.D. = 22.21, max = 100, min = 1	M = 52.05, S.D. = 21.17, max = 100, min = 0	M = 79.92, S.D. = 32.79, max = 100, min = 1
Interest (7-Likert scale: [1..7])	M = 5.14, S.D. = 1.65, max = 7, min = 1	M = 3.42, S.D. = 1.65, max = 6.25, min = 1	M = 4.40, S.D. = 1.96, max = 7, min = 1
Strength (overall) (5-Likert scale: [0.5..4.5])	M = 3.46, S.D. = 1.22, max = 4.5, min = 0.5	M = 3.20, S.D. = 1.31, max = 4.5, min = 0.5	M = 3.29, S.D. = 1.38, max = 4.5, min = 0.5
Strength (pro-tweet only)	M = 3.46, S.D. = 1.23 max = 4.5, min = 0.5	M = 3.50, S.D. = 1.21 max = 4.5, min = 0.5	M = 2.87, S.D. = 1.43 max = 4.5, min = 0.5
Strength (con-tweet only)	M = 3.47, S.D. = 1.22 max = 4.5, min = 0.5	M = 2.92, S.D. = 1.35 max = 4.5, min = 0.5	M = 3.73, S.D. = 1.18 max = 4.5, min = 0.5

for BRS. Cronbach's alpha of 0.731 and 0.819 showed acceptable and good consistency for NFC and BRS, respectively. Lastly, we obtained WPCS from the sum of the agreeing responses to each liberal policy and the disagreeing responses to each conservative response. Thus, positive WPCS indicates a higher inclination towards liberal ideologies, while negative WPCS indicates a higher inclination towards conservative ideologies. The distributions of each measure were as follows: CRT (M = 4.21, S.D. = 1.97, max = 7, min = 0), NFC (M = 3.56, S.D. = 0.42, max = 4.33, min = 2.50), BRS (M = 3.17, S.D. = 0.77, max = 4.60, min = 1), and WPCS (M = 9.33, S.D. = 6.32, max = 21, min = −5). We found that the distribution of WPCS was skewed towards strong liberal ideologies. The statistics suggest that 16 individuals scored above 10, 24 scored between 1 and 10, and 2 scored between -5 and 0. This implies that most of our participants held moderate liberal ideologies while the lower end of our population implied those who held either neutral or leaned slightly towards conservative ideologies.

3.4.2 Contextual Factors. We obtained the participants' topic interest levels, Interest, which showed good internal consistency among the 4-item questions (Cronbach's alpha = 0.87). In addition, we obtained the perceived issue strength score, Strength, of each tweet evaluated by the participant by taking the unsigned distance between the self-report tweet issue strength (TP: 10-Likert scale, from 1: strongly opposing to 10: strongly supporting) and its neutral absolute (score of 5.5), *i.e.*, Strength = |TP − 5.5|. Table 3 summarises the statistics of the topic interest levels and the stimulus perceived issue strength score for each topic reported by our participants.

3.4.3 Ideological Stance and Congruence. We derived the participant's ideological stance in two ways: **implicit** (Stance_Imp) collected from the word association test (7-Likert scale), and **explicit** (Stance_Exp) collected from a self-report question (continuous from 0 to 100). The statistics of the implicit and explicit stances are also shown in Table 3. We found that most of our participants rather expressed *pro* attitudes for abortion rights and same-sex marriage while showing neutral stances on the issue of vegetarianism.

Subsequently, we derived the ideological congruence score of the user stimulus from the product of the ideological stances of the user and the tweet. Denoting $\textsc{Stance}(T)$ as the ideological stance of tweet T (-1: supporting the *con* stance and +1: supporting the *pro* stance), we calculated the explicit congruence (Cong_Exp) and implicit congruence (Cong_Imp) between participant P and tweet T using the following equations.

$$\textsc{Cong\_Imp}(P,T) = \textsc{Stance\_Imp}(P) \times \textsc{Stance}(T)$$
$$\textsc{Cong\_Exp}(P,T) = (\textsc{Stance\_Exp}(P) - 50) \times \textsc{Stance}(T)$$

We cross-checked Cong_Imp with Cong_Exp and found a Pearson correlation of 0.946, indicating that they were highly correlated. This ensures our internal validity as participants' self-assessments aligned with the implicit measures. Therefore, we report our analysis using Cong_Imp, *i.e.*, the implicit measure for ideological congruence.

3.5 Measures

3.5.1 Information-Seeking Intention. We derived the order of each tweet headline directly from the participant's final ranking of eight headlines. Each headline was labelled between 1 and 8, where 1 and 8 represented the most desirable headline to read and the least desirable headline to read accordingly.

3.5.2 Information Recall. We assessed the recall ability from the written, recalled responses in the free-recall task. In particular, we rated how well the response matched the presented tweets. Two researchers independently coded each recalled response according to the content of each tweet mentioned. Subsequently, for each matched tweet, they individually gave a rating for the richness of the response on a 4-item Likert-style scale (0: little or no mention of the tweet, 1: somewhat rich, 2: moderately rich, and 3: very rich). In particular, we scored 3 for the recalled tweet if it stated all aspects and details and closely resembled the original tweet.

A score of 2 was given if the recalled tweet was incomplete but featured more than one aspect or detail. A score of 1 was given if the recall mentioned only one aspect of the original tweet. A score of 0 was given if the tweet was not mentioned in the recalled response. We achieved an inter-rater reliability (Cohen's Kappa) of 0.779, indicating good consistency between the two raters. In summary, for each tweet, we derived a measure for the recall ability as a floor average of the ratings from two raters, *i.e.*, $\texttt{Rating}(T) = \lfloor(\texttt{Rating}_{R1}(T) + \texttt{Rating}_{R2}(T))/2\rfloor$.

3.5.3 Information Interpretation. Among the three information interpretation rating items (5-Likert scale) collected in the individual tweet evaluation task, a Cronbach's alpha of 0.783 indicated acceptable internal consistency across the three items. Therefore, we used an average of the three items as our measure.

4 Results

Given that the derived measures are all ordinal data, we performed mixed-effect ordinal regression analyses using the Cumulative Linked Mixed Models (CLMM) [19] to assess the effect of the user-stimulus ideological congruence and individual and contextual factors. Furthermore, we examined the interplay of individual and contextual factors with the reliance on ideological congruence as a *heuristic* for confirmation bias. To do so, we assessed the interaction effects between the ideological congruence, Cong_Imp and one of the individual or contextual factors in three scenarios that aimed at stimulating confirmation bias [144]: headline ranking, news feed free-recall, and individual tweet evaluation.

We performed the regression analyses to examine the main and interaction effects between Cong_Imp and each of the predictors: (a) Interest, (b) CRT, (c) WPCS, (d) Strength, (e) NFC, and (f) BRS. To avoid colinearity issues, we ran the regression analysis separately for each measure and predictor. We also accounted for random effects from participants and stimuli to reflect our repeated-measure study design. Therefore, the formula for our CLMM models is `(measure ~ 1 + cong_imp * predictor + (1|participant_id) + (1|stimulus_id))`. We standardised all predictors before performing the regression analyses. From the data collected, we discarded 29 observations (2.87%) due to data losses (*e.g.*, the Qualtrics survey did not successfully capture some of the participants' responses). We note that our statistical models (CLMMs) are robust against missing data, which was minimal in our experiment.

Further, we looked into each interaction effect by performing posthoc regression analyses using CLMM to compare the estimated regression coefficient (β) or the main effect size of ideological congruence (Cong_Imp) between two different conditions of the interaction predictor variables, separated by its median value. Following the approach in Clogg et al. [22], we then performed a two-sampled Z-test to compare the regression coefficients (*i.e.*, the main effect size of Cong_Imp on the measurement variable) between two conditions (High and Low). The formula for the Z statistic is denoted in Equation 4, where β, σ, and n represent the estimated regression coefficient (main effect size), the standard error, and the sample size, respectively.

$$Z = \frac{\beta_{\text{High}} - \beta_{\text{Low}}}{\sqrt{\sigma^2_{\text{High}}/n_{\text{High}} + \sigma^2_{\text{Low}}/n_{\text{Low}}}}$$

In this section, we present the results of the regression and posthoc analyses, which reveal the relationships between the occurrences of confirmation bias and their influencing factors. Tables 4, 5, and 6 show a brief summary of regression results for information-seeking intention, recall, and interpretation, respectively. Table 7 depicts test statistics from the posthoc regression analyses and coefficient comparisons. We include the full regression tables in the supplementary materials.

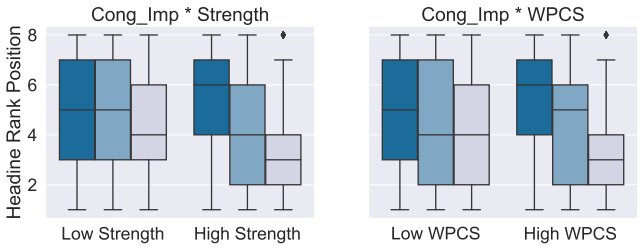
4.1 Information-Seeking Intention

4.1.1 Regression Analyses. We ran a mixed-effect ordinal regression with the ranking position of each tweet headline as the dependent variable on the data collected from the headline ranking task (Models 1a-1f). We found significant interaction effects of Cong_Imp × WPCS ($p < 0.001$, $\beta = 3.841$, $S.E. = 0.801$, $Z = 4.798$) and Cong_Imp × Strength ($p < 0.001$, $\beta = 1.334$, $S.E. = 0.581$, $Z = 2.295$). We did not detect significant interaction effects of CRT, NFC, BRS, and Interest with the implicit ideological congruence. Table 4 summarises the regression results for headline rank position. Figure 3a displays boxplots that illustrate the interaction effects of Cong_Imp × Strength and Cong_Imp × WPCS with x-axes being the interaction variables in two conditions, high (above or equal median) and low (below median).

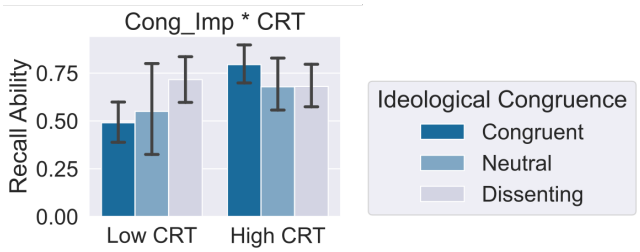
4.1.2 Posthoc Analyses. To closely examine the interaction effects, we performed posthoc comparisons of regression coefficients. On the interaction effect Cong_Imp × WPCS, we found that participants who held a relatively strong liberal ideology tended to rely more on ideological congruence than those more moderately oriented ($p < 0.001$, $Z = 52.076$; High WPCS: $p < 0.001$, $\beta = 2.618$, $S.E. = 0.431$, $Z = 6.073$; Low WPCS: $p < 0.001$, $\beta = 1.183$, $S.E. = 0.436$, $Z = 2.711$). Subsequently, posthoc analyses on the interaction effect Cong_Imp × Strength suggested that participants relied on ideological congruence for headlines of ideologically strong tweets more than those of ideologically neutral tweets ($p < 0.001$, $Z = 40.968$; High Strength: $p < 0.001$, $\beta = 2.305$, $S.E. = 0.408$, $Z = 5.698$; Low Strength: $p < 0.001$, $\beta = 1.196$, $S.E. = 0.408$, $Z = 2.931$).

4.2 Information Recall

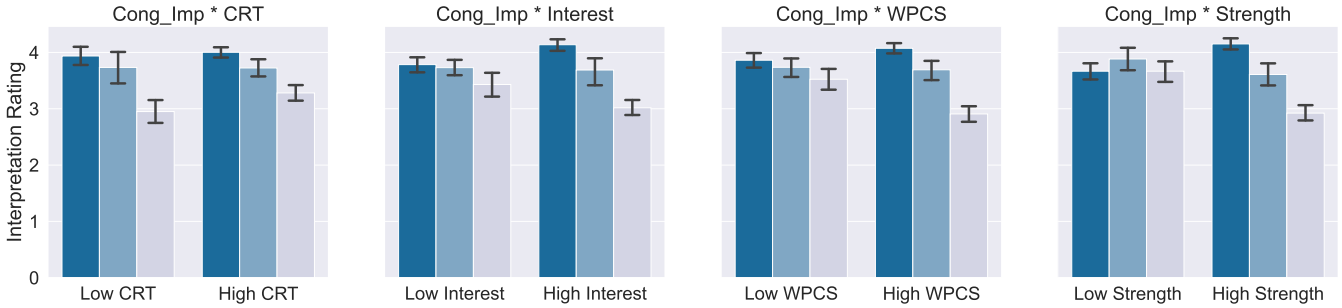
4.2.1 Regression Analyses. We performed a mixed-effect ordinal regression on the recall ability scores (Models 2a-2f). We only detected significant interaction effects of Cong_Imp × CRT ($p = 0.005$, $\beta = 0.018$, $S.E. = 0.006$, $Z = 2.781$). We did not detect significant interaction effects of NFC, BRS, Interest, and Strength with the implicit ideological congruence. Interestingly, we found a significant main effect of the tweet's issue strength on the recall ability ($p = 0.011$, $\beta = 0.628$, $S.E. = 0.245$, $Z = 2.560$). Specifically, we found individuals tended to recall the tweet better if they perceived it as ideologically stronger. Table 5 summarises the regression results for recall ability. Figure 3b illustrates the interaction effect of Cong_Imp × CRT.



(a) Interaction effects on the headline rank position. Posthoc regression analyses show that the effect of Cong_Imp on the headline rank position in High Strength and High WPCS conditions are stronger than those in Low Strength and Low WPCS, respectively.



(b) Interaction effects on the recall ability. Posthoc regression analyses show that the effect of Cong_Imp on the recall ability is significant in the Low CRT condition, but insignificant in the High CRT condition.



(c) Interaction effects on the information interpretation ratings. Posthoc regression analyses show that the effect of Cong_Imp on information interpretation ratings in High Interest, High WPCS, and Low CRT conditions are stronger than those in Low Interest, Low WPCS, and High CRT, respectively. The effect of Cong_Imp is significant in the High Strength condition, but insignificant in the Low Strength condition.

Figure 3: Interaction effects on (a) the headline rank position (Cong_Imp×Strength and Cong_Imp×WPCS), (b) the recall ability (Cong_Imp × CRT), and (c) the information interpretation ratings (Cong_Imp × CRT,Cong_Imp × Interest,Cong_Imp × WPCS, and Cong_Imp × Strength). In each interaction plot, different levels of Cong_Imp are compared: congruent ([3..7]), neutral ([-2..+2]), and dissenting ([-7..-3]).

4.2.2 Posthoc Analyses. From posthoc comparisons of regression coefficients, we found that individuals who scored lower on CRT tended to show better recall of ideologically dissenting tweets than congruent ones (Low CRT: $p = 0.022$, $\beta = -0.8073$, $S.E. = 0.354$, $Z = -2.280$). For individuals with higher CRT scores, however, we detected no significant linear relationship between the recall ability and the implicit ideological congruence (High CRT: *n.s.*, $\beta = 0.367$, $S.E. = 0.381$, $Z = 0.965$).

4.3 Information Interpretation

4.3.1 Regression Analyses. We performed mixed effect ordinal regression analyses on the information interpretation ratings (models 3a-3f). We found significant interaction effects of Cong_Imp × CRT ($p = 0.003$, $\beta = -1.745$, $S.E. = 0.5911$, $Z = -2.953$), Cong_Imp × WPCS ($p < 0.001$, $\beta = 5.581$, $S.E. = 0.854$, $Z = 6.529$), Cong_Imp × Interest ($p < 0.001$, $\beta = 2.4273$, $S.E. = 0.730$, $Z = 3.324$) and Cong_Imp × Strength ($p < 0.001$, $\beta = 4.930$, $S.E. = 0.646$, $Z = 7.632$). We found no significant interaction effect of NFC and BRS with the implicit ideological congruence. Table 6 summarises the regression results for information interpretation. Figure 3c shows barplots visualising the interaction effects of Cong_Imp × CRT, Cong_Imp × Interest, Cong_Imp × WPCS and Cong_Imp × Strength.

4.3.2 Posthoc Analyses. Posthoc comparisons revealed that, when interpreting information, participants who held a rather extreme ideology tended to rely on ideological congruence more than those with somewhat nuanced ideology ($p < 0.001, Z = 63.858$; High WPCS: $p < 0.001$, $\beta = 3.778$, $S.E. = 0.457$, $Z = 8.259$; Low WPCS: $p < 0.001$, $\beta = 1.910$, $S.E. = 0.524$, $Z = 3.645$). Individuals who exhibited higher topic interest also relied on ideological congruence more than those with lower interest levels ($p < 0.001$, $Z = 65.942$; High Interest: $p < 0.001$, $\beta = 3.302$, $S.E. = 0.439$, $Z = 7.517$; Low Interest: $p < 0.001$, $\beta = 1.450$, $S.E. = 0.620$, $Z = 2.337$). We found individuals who scored lower on CRT tended to exhibit a stronger effect size of ideological congruence than those who scored higher ($p < 0.001$, $Z = -4.055$; High CRT: $p < 0.001$, $\beta = 2.833$, $S.E. = 0.442$, $Z = 6.399$; Low CRT: $p < 0.001$, $\beta = 2.955$, $S.E. = 0.514$, $Z = 5.750$). However, for the issue strength, we only detected a significant effect size of Cong_Imp on information interpretation the high Strength group (High Strength: $p < 0.001$, $\beta = 3.894$, $S.E. = 0.408$, $Z = 9.533$), while the low Strength group showed no significant linear relationship (Low Strength: *n.s.*, $\beta = 0.915$, $S.E. = 0.547$, $Z = 1.673$). In other words, we found that individuals tended to rely on ideological congruence when interpreting tweets with a strong ideological stance.

Table 4: Summary of main and interaction effects between ideological congruence and bias susceptibility factors from the ordinal mixed-effect regression analysis on headline ranking (DV: Rank Position). ** marks a significant effect.

DV: Rank Position	Coef. (β)	S.E.	Z	p-value
Cong_Imp	1.097	0.662	1.656	0.097
Interest	−0.486	0.364	−1.335	0.182
Cong_Imp × Interest	1.110	0.637	1.743	0.081
Model Summary (1a)	Log-Likelihood = −1979.60, AIC = 3983.19, $\text{Cond}_H = 1.9 \times 10^3$			
Cong_Imp	2.438	0.427	5.702	$< 0.001^{**}$
CRT	0.356	0.343	1.040	0.299
Cong_Imp × CRT	−0.672	0.558	−1.202	0.229
Model Summary (1b)	Log-Likelihood = −1980.48, AIC = 3984.97, $\text{Cond}_H = 9.6 \times 10^2$			
Cong_Imp	1.038	0.549	1.889	0.059
Strength	−0.671	0.370	−1.814	0.070
Cong_Imp × Strength	1.334	0.581	2.295	0.022^{**}
Model Summary (1c)	Log-Likelihood = −1920.42, AIC = 3864.84, $\text{Cond}_H = 1.5 \times 10^3$			
Cong_Imp	−0.652	0.665	−0.980	0.327
WPCS	−1.822	0.469	−3.887	$< 0.001^{**}$
Cong_Imp × WPCS	3.841	0.801	4.798	$< 0.001^{**}$
Model Summary (1d)	Log-Likelihood = −1969.39, AIC = 3962.78, $\text{Cond}_H = 2.2 \times 10^3$			
Cong_Imp	1.499	0.591	2.535	0.011^{**}
NFC	−0.447	0.449	−0.995	0.320
Cong_Imp × NFC	0.893	0.738	1.210	0.226
Model Summary (1e)	Log-Likelihood = −1980.48, AIC = 3984.96, $\text{Cond}_H = 1.9 \times 10^3$			
Cong_Imp	2.332	0.505	4.614	$< 0.001^{**}$
BRS	0.138	0.442	0.314	0.753
Cong_Imp × BRS	−0.425	0.714	−0.595	0.552
Model Summary (1f)	Log-Likelihood = −1981.00, AIC = 3985.99, $\text{Cond}_H = 1.5 \times 10^3$			

Table 5: Summary of main and interaction effects between ideological congruence and bias susceptibility factors from the ordinal mixed-effect regression analysis on recall ability scores (DV: Recall Ability). ** marks a significant effect.

DV: Recall Ability	Coef. (β)	S.E.	Z	p-value
Cong_Imp	0.007	0.079	0.094	0.925
Interest	0.018	0.027	0.684	0.494
Cong_Imp × Interest	−0.001	0.006	−0.203	0.839
Model Summary (2a)	Log-Likelihood = −1012.46, AIC = 2040.93, $\text{Cond}_H = 9.7 \times 10^4$			
Cong_Imp	−0.076	0.034	−2.287	0.022^{**}
CRT	0.073	0.069	1.045	0.296
Cong_Imp × CRT	0.018	0.006	2.781	0.005^{**}
Model Summary (2b)	Log-Likelihood = −1008.30, AIC = 2032.60, $\text{Cond}_H = 1.9 \times 10^4$			
Cong_Imp	−0.050	0.045	−1.121	0.262
Strength	0.628	0.245	2.560	0.011^{**}
Cong_Imp × Strength	0.054	0.049	1.111	0.266
Model Summary (2c)	Log-Likelihood = −1008.59, AIC = 2033.19, $\text{Cond}_H = 7.7 \times 10^2$			
Cong_Imp	−0.018	0.040	−0.459	0.647
WPCS	0.035	0.021	1.675	0.093
Cong_Imp × WPCS	8×10^{-4}	0.003	0.319	0.750
Model Summary (2d)	Log-Likelihood = −1011.31, AIC = 2038.62, $\text{Cond}_H = 1.1 \times 10^5$			
Cong_Imp	−0.009	0.124	−0.078	0.938
NFC	0.028	0.055	0.513	0.608
Cong_Imp × NFC	7.638×10^{-5}	0.006	0.014	0.989
Model Summary (2e)	Log-Likelihood = −1012.59, AIC = 2041.17, $\text{Cond}_H = 4.3 \times 10^6$			
Cong_Imp	0.047	0.056	0.845	0.398
BRS	0.013	0.018	0.765	0.444
Cong_Imp × BRS	−0.001	0.001	−1.080	0.280
Model Summary (2f)	Log-Likelihood = −1011.84, AIC = 2039.69, $\text{Cond}_H = 2.4 \times 10^6$			

Table 6: Summary of main and interaction effects between ideological congruence and bias susceptibility factors from the ordinal mixed-effect regression analysis on individual tweet interpretation (DV: Info. Interpretation). ∗∗ marks a significant effect.

DV: Info. Interpretation	Coef. (β)	S.E.	Z	p-value
Cong_Imp	0.875	0.775	1.130	0.259
Interest	−1.404	0.591	−2.376	0.018^{**}
Cong_Imp × Interest	2.427	0.730	3.324	$< 0.001^{**}$
Model Summary (3a)	Log-Likelihood = −2098.41, AIC = 4230.82, $\text{Cond}_H = 4.2 \times 10^3$			
Cong_Imp	4.106	0.498	8.244	$< 0.001^{**}$
CRT	1.317	0.593	2.223	0.026^{**}
Cong_Imp × CRT	−1.745	0.591	−2.953	0.003^{**}
Model Summary (3b)	Log-Likelihood = −2099.26, AIC = 4232.52, $\text{Cond}_H = 2.8 \times 10^3$			
Cong_Imp	−0.529	0.599	−0.885	0.376
Strength	−2.104	0.410	−5.124	$< 0.001^{**}$
Cong_Imp × Strength	4.930	0.646	7.632	$< 0.001^{**}$
Model Summary (3c)	Log-Likelihood = −2072.17, AIC = 4178.35, $\text{Cond}_H = 2.5 \times 10^3$			
Cong_Imp	−0.857	0.714	−1.201	0.229
WPCS	−2.756	0.757	−3.642	$< 0.001^{**}$
Cong_Imp × WPCS	5.581	0.854	6.529	$< 0.001^{**}$
Model Summary (3d)	Log-Likelihood = −2082.53, AIC = 4199.05, $\text{Cond}_H = 4.5 \times 10^3$			
Cong_Imp	2.745	0.676	4.062	$< 0.001^{**}$
NFC	−0.357	0.776	−0.461	0.645
Cong_Imp × NFC	0.615	0.824	0.747	0.455
Model Summary (3e)	Log-Likelihood = −2103.76, AIC = 4241.51, $\text{Cond}_H = 4.7 \times 10^3$			
Cong_Imp	3.381	0.536	6.312	$< 0.001^{**}$
BRS	0.455	0.783	0.581	0.561
Cong_Imp × BRS	−0.419	0.721	−0.581	0.561
Model Summary (3f)	Log-Likelihood = −2103.81, AIC = 4241.62, $\text{Cond}_H = 4.3 \times 10^3$			

Table 7: Summary of posthoc comparison statistics for each of the interaction effects on three measures (DV: Rank Position, Recall Ability, and Information Interpretation). Inferential statistics for posthoc regression analyses (via CLMMs) and coefficient comparisons (via two-sampled Z-tests) are shown on the left and right, respectively. N represents the sample size, n_{px} denotes the number of participants included in the sample, and ∗∗ marks a significant effect.

DV	Condition	N (n_{px})	Coef. (β)	S.E.	Z	p-value	Coefficient Comparison
Rank Position	Strength (High)	634 (41)	2.305	0.405	5.698	$< 0.001^{**}$	High > Low
	Strength (Low)	345 (39)	1.196	0.408	2.931	0.003^{**}	$Z = 40.968, p < 0.001$
Rank Position	WPCS (High)	499 (21)	2.618	0.431	6.073	$< 0.001^{**}$	High > Low
	WPCS (Low)	480 (20)	1.183	0.436	2.711	0.007^{**}	$Z = 52.076, p < 0.001$
Recall Ability	CRT (High)	648 (24)	0.367	0.380	0.965	0.335	Not applicable
	CRT (Low)	331 (14)	−0.807	0.354	−2.280	0.022^{**}	
Info. Interpret.	CRT (High)	648 (27)	2.833	0.442	6.399	$< 0.001^{**}$	Low > High
	CRT (Low)	331 (14)	2.955	0.514	5.750	$< 0.001^{**}$	$Z = -4.055, p < 0.001$
Info. Interpret.	Strength (High)	634 (41)	3.894	0.408	9.533	$< 0.001^{**}$	Not Applicable
	Strength (Low)	345 (39)	0.915	0.547	1.673	0.094	
Info. Interpret.	Interest (High)	499 (21)	3.302	0.439	7.517	$< 0.001^{**}$	High > Low
	Interest (Low)	480 (20)	1.450	0.620	2.337	0.0194^{**}	$Z = 65.942, p < 0.001$
Info. Interpret.	WPCS (High)	499 (21)	3.778	0.457	8.259	$< 0.001^{**}$	High > Low
	WPCS (Low)	480 (20)	1.910	0.524	3.645	$< 0.001^{**}$	$Z = 63.858, p < 0.001$

5 Discussion

In this study, we investigated the roles of different individual and contextual factors in amplifying and moderating confirmation biases. Through three task scenarios where confirmation bias generally manifests, we exposed participants to tweet-like content items that stated a strong opinion on controversial topics. With the individual and contextual differences we collected in the study, we found that the tendency for effortful thinking, strong political beliefs, and the perceived issue strength of the tweet influence the occurrences of confirmation bias. In the remainder of this section, we discuss our findings in detail, followed by the practical and ethical implications for designing and tailoring context-aware interventions to effectively mitigate cognitive biases.

5.1 Amplifiers for Confirmation Bias

5.1.1 Content's Perceived Issue Strength. We found that the perceived tweet's issue strength interacted with the effects of ideological congruency on the information-seeking intention and information interpretation ratings. In both tasks, the effect of ideological congruency is significant when individuals interact with tweets perceived as ideologically strong, with the headline ranking task also showing that the effect of ideological congruency is stronger when the participants perceived that the tweet's issue strength was stronger. This finding implies that content perceived as a strong issue may be more likely to trigger confirmation bias. It also resonates with Zhao et al. [150], who suggested that users were more likely to share online health articles with a strong opinion stance. In addition, we found that individuals tended to recall better tweets that were perceived as ideologically stronger. This finding aligns with previous research on human memory [28, 72], showing that people tend to remember better emotionally valenced stimuli than neutral stimuli. Our results provide an empirical contribution to the confirmation bias literature as we shed light on the role of content's perceived issue strength on memory recall.

5.1.2 Individual's Political Attitudes. We found that a strong leaning towards Liberalism, reflected through high WPCS scores in our population, amplified the effects of confirmation bias on information-seeking intention and information interpretation. Notably, individuals with relatively strong liberal beliefs tended to rank higher (*i.e.*, feel inclined to read the entire tweet) headlines of tweets deemed congruent with their beliefs. They perceived it differently from those ideologically dissenting. In other words, this suggests that individuals with relatively strong political leanings may be more susceptible to using mental shortcuts and cognitive biases. Prior studies also support our findings; for example, Pennycook and Rand [103], as well as Traberg and van der Linden [139], suggested that political partisanship affects how individuals evaluate information as they perceive news with politically opposing stances or sources as less reliable. Therefore, our findings extend the prior literature: we demonstrate that individuals with strong liberal leanings are more susceptible to confirmation bias than those with neutral (or moderate conservative) beliefs.

5.1.3 Individual's Thinking Styles. Our results also highlighted that individuals with a lower tendency for effortful thinking, *i.e.*, those who scored lower on the Cognitive Reflection Test (CRT), relied significantly on ideological congruence when interpreting and recalling information. In the information interpretation scenario, we found a stronger effect size of ideological congruence when comparing individuals who scored lower on CRT and those who scored higher. Importantly, we found that the tendency for low effortful thinking determined how information is recalled: individuals with a lower effortful thinking tendency recalled better information that opposed their beliefs. This aligns with findings from Greene et al. [50], who reported that individuals with a lower effortful thinking tendency, measured similarly through CRT, formed more false memories than those with a higher tendency. However, our results contrast with Strømsø et al. [130], who found that individuals who better recalled belief-inconsistent information tended to score higher on CRT. While our work quantified the recall ability on a 4-Likert scale, they measured recall using a binary construct (*i.e.*, whether the recalled response is consistent with the original content or not). More research is needed to investigate the joint role of effortful thinking and prior beliefs in information recall.

5.1.4 Task Design and Modality. When considering all scenarios in conjunction, the factors we identified in this study amplify confirmation bias differently in each task. For example, while we found that the tweet's issue strength and the individual's effortful thinking tendency influence confirmation bias in information-seeking intention, the former did not show an interaction effect on recall ability. Similarly, the individual's topic interest only appeared as a susceptibility factor of confirmation bias in the information interpretation scenario. This finding, therefore, suggests that the *interaction context*, *i.e.*, the nature of the task, could be an influencing factor to bias susceptibility. Similarly, Vedejová and Čavojová [144] investigated confirmation bias across three scenarios (information seeking, interpretation, and recall) but did not find an effect of confirmation bias in information recall, in which the authors acknowledged that the nature of the task deployed in the study could influence how confirmation bias manifests. In the context of AI trust calibration, Ha and Kim [53] showed that different modalities of interventions (visual and textual explanation) could influence the effectiveness of confirmation bias mitigation. In psychology, Jonas et al. [65] suggested that the more natural setting of the information task leads to a stronger biased information processing, and, therefore, the choice of experimental design could influence (or confound) the occurrences of confirmation bias. In this study, however, we did not investigate the interaction effects between the contextual and individual factors on confirmation bias (*e.g.*, would effortful thinking tendency interact with the choice of task design?). Thus, we call for future research to consider multiple factors in conjunction when studying bias susceptibility.

5.2 Practical and Ethical Implications for Context-Aware Intervention Design

Synthesising these insights, our results can help inform the design of cognitive bias interventions by taking into account bias susceptibility factors, namely, the content's perceived issue strength, the user's political leaning, and thinking styles. With ideologically strong information amplifying confirmation bias, our findings suggest platforms could target content items that tend to be perceived by users as a strong stance. Consequently, platforms could soften

the issue strength of social media content using linguistic models [78, 107, 138] to detect and adjust the content's stance and sentiment towards a more nuanced perspective to help safeguard users from falling victim to their biases. It is worth noting that users may perceive the same content differently, potentially seeing it as stronger than intended by the content creator. On the other hand, intervention designers could statistically model how users perceive the issue strength of different expressions based on their rating of the content – the same measure we employed in this study (Strength) – and personalise interventions that adapt according to the content's tendency to be perceived as a firm stance. However, we acknowledge that the strength-softening measure may be viewed as a form of censorship and benevolent paternalism (*i.e.*, limiting the agency and controlling the content's stance in the *best interest* of the people). This concern is similar to the critiques about nudging as it limits users' autonomy over their decision-making [64, 132]. It may introduce novel production incentives as well as externalities [71] as content creators may divert from producing content with the potential to be demoted on the platform. On the other hand, we argue that platforms carefully consider their measures while giving users the autonomy to decide which version of the content – original or filtered – they would like to see.

Furthermore, platforms could specifically consider users' individual differences. Preventive interventions, such as psychological inoculation [84], media literacy building [52], or imposing safeguarding mechanism [32, 82], could also be targetted to specific user groups to train and fortify them against manipulation. The abovementioned bias susceptibility factors can be inferred via users' daily usage and interaction on social media platforms [100] or collected via one-off questionnaires. Nevertheless, we note that with the ability to identify individuals' bias susceptibility, platforms should leverage this data ethically not to amplify their users' cognitive biases. Importantly, by identifying biases, there is a risk of abuse to reaffirm and influence people's beliefs and decision-making. The Cambridge Analytica scandal [16, 60] is a prominent example where intimate knowledge about users' psychological traits was used to tailor targeted messaging to manipulate their opinion formation and decision-making. Therefore, the ability to determine individuals' specific tendencies and bias susceptibility should be treated with great caution. We envision that platforms could practically provide transparency of the personalised interventions through informed consent, giving users the awareness of what data are being collected, as well as an explanation of how and why a certain intervention is being tailored to them [152]. Users should always be given the autonomy to review what interventions are being applied to them, the potential impacts for them (*e.g.*, this intervention may subconsciously steer your news feed behaviour), and the ability to opt-out. We also acknowledge that data protection laws (for example, European Union's GDPR[3]) may restrict the ability to collect sensitive data which are used to inform personalised interventions.

In summary, our work paves the way towards *context-aware* interventions which adapt to the *user* and *interaction context*. The literature clearly indicates that there is no one-size-fits-all approach to mitigating cognitive biases because these cognitive tendencies manifest differently depending on the context of users, systems, and their interactions [2, 91, 110]. Cresci et al. [24] also argued that interventions could be shifted from a platform-centred approach to a personalised manner through user and context modelling. By extending the notion of context-awareness [116], we can develop computing systems that personalise not only to one's cognitive biases but, at the same time, mitigate their adverse effects using the same characteristics deduced from users' interaction data [97]. Nonetheless, it is unclear how effective personalised, context-aware interventions would be. Rieger et al. [113] investigated the effect of cognitive reflection style (collected through CRT) on the effectiveness of nudging and boosting interventions. While they did not find a significant effect, the authors argued that the effect might have been moderated by other individual and contextual factors. In line with their work, we envision that future research evaluates the effectiveness of personalised cognitive bias interventions across different populations, contexts, and task designs.

[3]https://gdpr-info.eu/issues/personal-data/

5.3 Limitations

There are several limitations to our study. First, the distribution of WPCS indicated that most of our participants reported strong liberal ideologies, while those who held conservative beliefs were small in number. This is a common phenomenon when recruiting study participants, especially from a university campus [90]. This population also tends to have higher education levels and more developed critical thinking abilities, as shown in our study participants' demographics and distribution of CRT scores. While the finding indicates that the tendency for political beliefs was an influencing factor of confirmation biases, it only gives a one-sided picture as we could only draw a comparison between strong liberals on one end and neutrals or moderate conservatives on the other end. The literature suggests that individuals with conservative beliefs could be more susceptible to misinformation and to using mental shortcuts [34, 51, 67]. Meanwhile, Ditto et al. [27] and Enders et al. [37] argued that, in the US political context, liberals are not less susceptible than conservatives. Therefore, collecting more participant samples from the conservative end would give a more complete picture of bias susceptibility, allowing a better generalisability of our findings. We also acknowledge that the number of participants in this study is limited due to its in-person setting. While our sample size (N = 42) is properly powered, we suggest future works could replicate our study through online experiments and recruit a larger, more heterogeneous set of participants.

We deployed our stimuli and tasks on Qualtrics. We acknowledge that it may not represent realistic information consumption scenarios on social media, which may consist of source cues, visual information, and social interaction with other users. Nevertheless, our main focus is on how the information is processed and how individual and contextual factors influence confirmation bias in such activities. Qualtrics, therefore, allowed us to separate confounds and closer study factors for bias susceptibility. We envision that future research further investigates bias susceptibility in higher fidelity settings, resembling real-world information consumption scenarios, while accounting for potential confounds, such as the effects of source bias and the coherence between prior beliefs and the stance of the information content [135].

We employed self-report questionnaires to gauge the tweet's perceived issue strength (Strength). This measure may be prone to subjectivity and, therefore, may not offer the best indicator for content's strong stance as applied in context-aware interventions. We recommend that future research consider using a crowd of people (that represent diverse political inclinations) to determine the issue strength of the media content.

Moreover, the recall responses and ability scores may be prone to noise. While we instructed participants to write down everything they could remember about the tweets, each participant approached this task differently. For example, some participants reported that they tried to summarise and elaborate the tweets into *pro* and *con* sides. At the same time, some listed the details of each tweet they remembered. Because the task did not restrict what the participants could write down, the richness of the recall responses depended on the participant's discretion. Some participants provided extensive recall, while some wrote only the critical aspects of each tweet. In future work, we suggest that free-recall measures could be accompanied by cued recall, *e.g.*, asking participants multiple-choice questions about the tweets or recognition tasks, where participants are asked to label items they remember.

Lastly, we used the word association test to gauge participants' ideological stances. Although it showed a strong correlation with the explicit self-assessment stance, Cong_Exp, our implicit measure may be prone to the same self-presentation [129], preference falsification issues [79], and partisan bias [13]. We suggest that future research consider measurements that better separate out these confounds, for example, the implicit association test [26, 124], to capture the nuanced strength of ideological stance and political concordances.

6 Conclusion

Cognitive biases offer useful heuristics that allow us to sift through the sheer amounts of online information quickly and effectively. At the same time, this comes at the cost of undermining the quality of our decision-making. Cognitive biases tend to be difficult to be effectively mitigated. People seem to be susceptible to acting on their biases to different degrees. In this work, we shed light on the influencing role of individual and contextual factors of cognitive biases in three scenarios: information-seeking intention, recall, and interpretation – three tasks commonly found when sifting through information online. Specifically, we investigated how these factors amplify confirmation bias – the reliance on prior beliefs – when exposed to ideologically polarised content. We found that the individual's strong political beliefs, low-effortful thinking tendency, and interest in the issue, as well as the content's perceived firm stance and the nature of the interaction with information, render users especially susceptible to confirmation bias. These insights pave the way towards more targeted safeguarding mechanisms and designing more effective, context-aware intervention systems that consider individual and contextual differences to mitigate cognitive biases, keep people safe online, and support more informed decision-making. Our findings inform measures on social media platforms to (1) reduce language that tends to be perceived as emotional or firm expressions and (2) target preventive interventions, such as safeguarding and media literacy-building mechanisms, on users with tendencies for low-effortful thinking and strong political beliefs. At the same time, designers should take these characteristics with great care and transparency, as they could open doors for paternalism and manipulation.

Acknowledgments

We thank participants in our studies and members of the HCI group at the University of Melbourne for their feedback, which helped to positively shape this paper.

References

[1] Alberto Acerbi. 2019. Cognitive attraction and online misinformation. *Palgrave Communications* 5, 1 (2019).

[2] Zhila Aghajari, Eric P. S. Baumer, and Dominic DiFranzo. 2023. Reviewing Interventions to Address Misinformation: The Need to Expand Our Vision Beyond an Individualistic Focus. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 87 (apr 2023), 34 pages. https://doi.org/10.1145/3579520

[3] Faisal Alatawi, Lu Cheng, Anique Tahir, Mansooreh Karami, Bohan Jiang, Tyler Black, and Huan Liu. 2021. A Survey on Echo Chambers on Social Media: Description, Detection and Mitigation. *arXiv preprint arXiv:2112.05084* (2021). https://arxiv.org/abs/2112.05084

[4] Jennifer Allen, Cameron Martel, and David G Rand. 2022. Birds of a Feather Don't Fact-Check Each Other: Partisanship and the Evaluation of News in Twitter's Birdwatch Crowdsourced Fact-Checking Program. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 245, 19 pages. https://doi.org/10.1145/3491102.3502040

[5] Leif Azzopardi. 2021. Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (Canberra ACT, Australia) *(CHIIR '21)*. Association for Computing Machinery, New York, NY, USA, 27–37. https://doi.org/10.1145/3406522.3446023

[6] William Badke. 2018. Fake news, confirmation bias, the search for truth, and the theology student. *Theological Librarianship* 11, 2 (2018), 4–7. https://doi.org/10.31046/tl.v11i2.519

[7] Ricardo Baeza-Yates. 2018. Bias on the Web. *Commun. ACM* 61, 6 (May 2018), 54–61. https://doi.org/10.1145/3209581

[8] Bence Bago, David G Rand, and Gordon Pennycook. 2020. Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of experimental psychology: general* 149, 8 (2020), 1608.

[9] Alexander Benlian. 2015. Web Personalization Cues and Their Differential Effects on User Assessments of Website Value. *Journal of Management Information Systems* 32, 1 (2015), 225–260. https://doi.org/10.1080/07421222.2015.1029394

[10] H. Berghel. 2018. Malice Domestic: The Cambridge Analytica Dystopia. *Computer* 51, 05 (may 2018), 84–89. https://doi.org/10.1109/MC.2018.2381135

[11] Nattapat Boonprakong, Xiuge Chen, Catherine Davey, Benjamin Tag, and Tilman Dingler. 2023. Bias-Aware Systems: Exploring Indicators for the Occurrences of Cognitive Biases When Facing Different Opinions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 27, 19 pages. https://doi.org/10.1145/3544548.3580917

[12] Nattapat Boonprakong, Benjamin Tag, and Tilman Dingler. 2023. Designing Technologies to Support Critical Thinking in an Age of Misinformation. *IEEE Pervasive Computing* (2023), 1–10. https://doi.org/10.1109/MPRV.2023.3275514

[13] John G. Bullock and Gabriel Lenz. 2019. Partisan Bias in Surveys. *Annual Review of Political Science* 22, Volume 22, 2019 (2019), 325–342. https://doi.org/10.1146/annurev-polisci-051117-050904

[14] Pavlo Burda, Luca Allodi, and Nicola Zannone. 2024. Cognition in Social Engineering Empirical Research: A Systematic Literature Review. *ACM Trans. Comput.-Hum. Interact.* 31, 2 (2024). https://doi.org/10.1145/3635149

[15] John T Cacioppo and Richard E Petty. 1982. The Need for Cognition. *Journal of personality and social psychology* 42, 1 (1982), 116.

[16] Carole Cadwalladr. 2017. The great British Brexit robbery: how our democracy was hijacked. *The Guardian* 7 (2017).

[17] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300733

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

[19] Rune Haubo B Christensen. 2015. A Tutorial on fitting Cumulative Link Models with the ordinal Package. *Retrieved from www.cran.r-project.org/package=ordinal* (2015).

[20] Russ Clay, Jessica M. Barber, and Natalie J. Shook. 2013. Techniques for Measuring Selective Exposure: A Critical Review. *Communication Methods and Measures* 7, 3-4 (2013), 147–171. https://doi.org/10.1080/19312458.2013.813925 arXiv:https://doi.org/10.1080/19312458.2013.813925

[21] Katherine Clayton, Jase Davis, Kristen Hinckley, and Yusaku Horiuchi. 2019. Partisan motivated reasoning and misinformation in the media: Is news from ideologically uncongenial sources more suspicious? *Japanese Journal of Political Science* 20, 3 (2019), 129–142. https://doi.org/10.1017/S1468109919000082

[22] Clifford C. Clogg, Eva Petkova, and Adamantios Haritou. 1995. Statistical Methods for Comparing Regression Coefficients Between Models. *Amer. J. Sociology* 100, 5 (1995), 1261–1293. https://doi.org/10.1086/230638 arXiv:https://doi.org/10.1086/230638

[23] Sonia Correa. 2003. Abortion is a global political issue. In *A DAWN Supplement for the World Social Forum, Porto Alegre, 23–28 January 2003.*

[24] Stefano Cresci, Amaury Trujillo, and Tiziano Fagni. 2022. Personalized Interventions for Online Moderation. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media* (Barcelona, Spain) *(HT '22)*. Association for Computing Machinery, New York, NY, USA, 248–251. https://doi.org/10.1145/3511095.3536369

[25] Kristen D. Deppe, Frank J. Gonzalez, Jayme L. Neiman, Carly Jacobs, Jackson Pahlke, Kevin B. Smith, and John R. Hibbing. 2015. Reflective liberals and intuitive conservatives: A look at the Cognitive Reflection Test and ideology. *Judgment and Decision Making* 10, 4 (2015), 314–331. https://doi.org/10.1017/S1930297500005131

[26] Tilman Dingler, Benjamin Tag, David A. Eccles, Niels van Berkel, and Vassilis Kostakos. 2022. Method for Appropriating the Brief Implicit Association Test to Elicit Biases in Users. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 243, 16 pages. https://doi.org/10.1145/3491102.3517570

[27] Peter H. Ditto, Brittany S. Liu, Cory J. Clark, Sean P. Wojcik, Eric E. Chen, Rebecca H. Grady, Jared B. Celniker, and Joanne F. Zinger. 2019. At Least Bias Is Bipartisan: A Meta-Analytic Comparison of Partisan Bias in Liberals and Conservatives. *Perspectives on Psychological Science* 14, 2 (2019), 273–291. https://doi.org/10.1177/1745691617746796 arXiv:https://doi.org/10.1177/1745691617746796 PMID: 29851554.

[28] Florin Dolcos, Kevin S LaBar, and Roberto Cabeza. 2006. The memory enhancing effect of emotion: Functional neuroimaging evidence. (2006).

[29] Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. 2021. This Is Not What We Ordered: Exploring Why Biased Search Result Rankings Affect User Attitudes on Debated Topics. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3404835.3462851

[30] James N. Druckman. 2012. The Politics of Motivation. *Critical Review: A Journal of Politics and Society* 24, 2 (2012), 199–216. https://doi.org/10.1080/08913811.2012.711022

[31] AH Eagly, S Chen, S Chaiken, and K Shaw-Barnes. 1999. The impact of attitudes on memory: An affair to remember. *PSYCHOLOGICAL BULLETIN* 125, 1 (JAN 1999), 64–89. https://doi.org/10.1037/0033-2909.125.1.64

[32] David A Eccles, Sherah Kurnia, Tilman Dingler, and Nicholas Geard. 2021. Three Preventative Interventions to Address the Fake News Phenomenon on Social Media. In *ACIS 2021 Proceedings*, Vol. 51. https://aisel.aisnet.org/acis2021/51

[33] Ullrich KH Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K Fazio, Nadia Brashier, Panayiota Kendeou, Emily K Vraga, and Michelle A Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology* 1, 1 (2022), 13–29.

[34] Ullrich K. H. Ecker and Li Chang Ang. 2019. Political Attitudes and the Processing of Misinformation Corrections. *Political Psychology* 40, 2 (2019), 241–260. https://doi.org/10.1111/pops.12494

[35] Ullrich K. H. Ecker, Stephan Lewandowsky, Olivia Fenton, and Kelsey Martin. 2014. Do people keep believing because they want to? Preexisting attitudes and the continued influence of misinformation. *Memory & Cognition* 42, 2 (Feb. 2014), 292–304. https://doi.org/10.3758/s13421-013-0358-x

[36] Scott Eidelman, Christian S. Crandall, Jeffrey A. Goodman, and John C. Blanchar. 2012. Low-Effort Thought Promotes Political Conservatism. *Personality and Social Psychology Bulletin* 38, 6 (2012), 808–820. https://doi.org/10.1177/0146167212439213 arXiv:https://doi.org/10.1177/0146167212439213 PMID: 22427384.

[37] Adam Enders, Christina Farhart, Joanne Miller, Joseph Uscinski, Kyle Saunders, and Hugo Drochon. 2023. Are Republicans and Conservatives More Likely to Believe Conspiracy Theories? *Political Behavior* 45, 4 (Dec. 2023), 2001–2024. https://doi.org/10.1007/s11109-022-09812-3

[38] Jonathan St BT Evans. 1989. *Bias in human reasoning: Causes and consequences.* Lawrence Erlbaum Associates, Inc.

[39] Jonathan St. B. T. Evans. 2008. Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology* 59, 1 (2008), 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629 arXiv:https://doi.org/10.1146/annurev.psych.59.103006.093629 PMID: 18154502.

[40] Jonathan St. B. T. Evans, Simon J. Handley, Helen Neilens, and David Over. 2010. The influence of cognitive ability and instructional set on causal conditional inference. *Quarterly Journal of Experimental Psychology* 63, 5 (2010), 892–909. https://doi.org/10.1080/17470210903111821 arXiv:https://doi.org/10.1080/17470210903111821 PMID: 19728225.

[41] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160.

[42] Leon Festinger. 1962. Cognitive Dissonance. *Scientific American* 207, 4 (1962), 93–106. http://www.jstor.org/stable/24936719

[43] Peter Fischer, Eva Jonas, Dieter Frey, and Stefan Schulz-Hardt. 2005. Selective exposure to information: The impact of information limits. *European Journal of social psychology* 35, 4 (2005), 469–492.

[44] Shane Frederick. 2005. Cognitive Reflection and Decision Making. *Journal of Economic Perspectives* 19, 4 (December 2005), 25–42. https://doi.org/10.1257/089533005775196732

[45] Peter Frost, Bridgette Casey, Kaydee Griffin, Luis Raymundo, Christopher Farrell, and Ryan Carrigan. 2015. The Influence of Confirmation Bias on Memory and Source Monitoring. *The Journal of General Psychology* 142, 4 (2015), 238–252. https://doi.org/10.1080/00221309.2015.1084987 arXiv:https://doi.org/10.1080/00221309.2015.1084987 PMID: 26649923.

[46] R. Kelly Garrett. 2009. Politically Motivated Reinforcement Seeking: Reframing the Selective Exposure Debate. *Journal of Communication* 59, 4 (12 2009), 676–699. https://doi.org/10.1111/j.1460-2466.2009.01452.x arXiv:https://academic.oup.com/joc/article-pdf/59/4/676/22323928/jjnlcom0676.pdf

[47] Christine Geeng, Savanna Yee, and Franziska Roesner. 2020. Fake News on Facebook and Twitter: Investigating How People (Don't) Investigate. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376784

[48] Gerd Gigerenzer. 2004. *Fast and Frugal Heuristics: The Tools of Bounded Rationality.* John Wiley & Sons, Ltd, Chapter 4, 62–88. https://doi.org/10.1002/9780470752937.ch4 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470752937.ch4

[49] Eduardo Graells-Garrido, Mounia Lalmas, and Ricardo Baeza-Yates. 2016. Data Portraits and Intermediary Topics: Encouraging Exploration of Politically Diverse Profiles. In *Proceedings of the 21st International Conference on Intelligent User Interfaces* (Sonoma, California, USA) *(IUI '16)*. Association for Computing Machinery, New York, NY, USA, 228–240. https://doi.org/10.1145/2856767.2856776

[50] Ciara M. Greene, Robert A. Nash, and Gillian Murphy. 2021. Misremembering Brexit: partisan bias and individual predictors of false memories for fake news stories among Brexit voters. *Memory* 29, 5 (2021), 587–604. https://doi.org/10.1080/09658211.2021.1923754 arXiv:https://doi.org/10.1080/09658211.2021.1923754 PMID: 33971789.

[51] Andrew Guess, Jonathan Nagler, and Joshua Tucker. 2019. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances* 5, 1 (2019), eaau4586. https://doi.org/10.1126/sciadv.aau4586 arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.aau4586

[52] Andrew M. Guess, Michael Lerner, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences* 117, 27 (2020), 15536–15545. https://doi.org/10.1073/pnas.1920498117 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1920498117

[53] Taehyun Ha and Sangyeon Kim. 2023. Improving Trust in AI with Mitigating Confirmation Bias: Effects of Explanation Type and Debiasing Strategy for Decision-Making with Explainable AI. *International Journal of Human–Computer Interaction* 0, 0 (2023), 1–12. https://doi.org/10.1080/10447318.2023.2285640 arXiv:https://doi.org/10.1080/10447318.2023.2285640

[54] Ulrike Hahn and Adam J.L. Harris. 2014. Chapter Two - What Does It Mean to be Biased: Motivated Reasoning and Rationality. Psychology of Learning and Motivation, Vol. 61. Academic Press, 41–102. https://doi.org/10.1016/B978-0-12-800283-4.00002-2

[55] Martie G. Haselton, Gregory A. Bryant, Andreas Wilke, David A. Frederick, Andrew Galperin, Willem E. Frankenhuis, and Tyler Moore. 2009. Adaptive Rationality: An Evolutionary Perspective on Cognitive Bias. *Social Cognition* 27, 5 (2009), 733–763. https://doi.org/10.1521/soco.2009.27.5.733 arXiv:https://doi.org/10.1521/soco.2009.27.5.733

[56] Martie G Haselton, Daniel Nettle, and Damian R Murray. 2015. The evolution of cognitive bias. *The handbook of evolutionary psychology* (2015), 1–20.

[57] Dallas Havens. 2022. *A Quick Look: The Debate Surrounding Ethical Vegetarianism.* https://dallashavens.wordpress.com/2022/07/08/a-quick-look-the-debate-

surrounding-ethical-vegetarianism/

[58] J.P. Henningham. 1996. A 12-item scale of social conservatism. *Personality and Individual Differences* 20, 4 (1996), 517–519. https://doi.org/10.1016/0191-8869(95)00192-1

[59] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. 2020. Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 48 (may 2020), 27 pages. https://doi.org/10.1145/3392854

[60] Jim Isaak and Mina J. Hanna. 2018. User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer* 51, 8 (2018), 56–59. https://doi.org/10.1109/MC.2018.3191268

[61] Kaixin Ji. 2023. Quantifying and Measuring Confirmation Bias in Information Retrieval Using Sensors. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing* (Cancun, Quintana Roo, Mexico) *(UbiComp/ISWC '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 236–240. https://doi.org/10.1145/3594739.3610765

[62] Ángel V Jiménez, Alex Mesoudi, and Jamshid J Tehrani. 2020. No evidence that omission and confirmation biases affect the perception and recall of vaccine-related information. *PloS one* 15, 3 (2020), e0228898. https://doi.org/10.1371/journal.pone.0228898

[63] Blair T Johnson and Alice H Eagly. 1989. Effects of involvement on persuasion: A meta-analysis. *Psychological bulletin* 106, 2 (1989), 290.

[64] Christine Jolls and Cass R. Sunstein. 2006. Debiasing through Law. *The Journal of Legal Studies* 35, 1 (2006), 199–242. https://doi.org/10.1086/500096 arXiv:https://doi.org/10.1086/500096

[65] Eva Jonas, Stefan Schulz-Hardt, Dieter Frey, and Norman Thelen. 2001. Confirmation bias in sequential information search after preliminary decisions: an expansion of dissonance theoretical research on selective exposure to information. *Journal of personality and social psychology* 80, 4 (2001), 557.

[66] Kristyn A. Jones, William E. Crozier, and Deryn Strange. 2017. Believing is Seeing: Biased Viewing of Body-Worn Camera Footage. *Journal of Applied Research in Memory and Cognition* 6, 4 (2017), 460–474. https://doi.org/10.1016/j.jarmac.2017.07.007

[67] John T Jost, Sander van der Linden, Costas Panagopoulos, and Curtis D Hardin. 2018. Ideological asymmetries in conformity, desire for shared reality, and the spread of misinformation. *Current Opinion in Psychology* 23 (2018), 77–83. https://doi.org/10.1016/j.copsyc.2018.01.003 Shared Reality.

[68] Dan M Kahan. 2015. The politically motivated reasoning paradigm. *Emerging Trends in Social & Behavioral Sciences, Forthcoming* (2015).

[69] Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Macmillan.

[70] Robert A. Kaufman, Michael Robert Haupt, and Steven P. Dow. 2022. Who's in the Crowd Matters: Cognitive Factors and Beliefs Predict Misinformation Assessment Accuracy. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 553 (nov 2022), 18 pages. https://doi.org/10.1145/3555611

[71] Devansh Kaushik. 2024. Policy Responses To Fake News On Social Media Platforms: A Law And Economics Analysis. *Statute Law Review* 45, 1 (02 2024), hmae013. https://doi.org/10.1093/slr/hmae013 arXiv:https://academic.oup.com/slr/article-pdf/45/1/hmae013/56770687/hmae013.pdf

[72] Elizabeth A Kensinger and Suzanne Corkin. 2003. Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words? *Memory & cognition* 31, 8 (2003), 1169–1180.

[73] Paul A. Klaczynski. 2000. Motivated Scientific Reasoning Biases, Epistemological Beliefs, and Theory Polarization: A Two-Process Approach to Adolescent Cognition. *Child Development* 71, 5 (2000), 1347–1366. https://doi.org/10.1111/1467-8624.00232

[74] Joseph T Klapper. 1960. The effects of mass communication. (1960).

[75] Jan Kleinnijenhuis. 2008. *Negativity*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781405186407.wbiecn005

[76] Keiichi Kobayashi. 2010. Strategic Use of Multiple Texts for the Evaluation of Arguments. *Reading Psychology* 31, 2 (2010), 121–149. https://doi.org/10.1080/02702710902754192 arXiv:https://doi.org/10.1080/02702710902754192

[77] Jon A Krosnick, Charles M Judd, and Bernd Wittenbrink. 2005. The measurement of attitudes. *The handbook of attitudes* 21 (2005), 76.

[78] Dilek Küçük and Fazli Can. 2020. Stance Detection: A Survey. *ACM Comput. Surv.* 53, 1, Article 12 (feb 2020), 37 pages. https://doi.org/10.1145/3369026

[79] Timur Kuran. 1997. *Private truths, public lies: The social consequences of preference falsification*. Harvard University Press.

[80] David Lazer, Matthew Baum, Nir Grinberg, Lisa Friedland, Kenneth Joseph, Will Hobbs, and Carolina Mattsson. 2017. Combating fake news: An agenda for research and action. (2017).

[81] Colin Lescarret, Valérie Le Floch, Jean-Christophe Sakdavong, Jean-Michel Boucheix, André Tricot, and Franck Amadieu. 2023. The impact of students' prior attitude on the processing of conflicting videos: a comparison between middle-school and undergraduate students. *European Journal of Psychology of Education* 38, 2 (June 2023), 519–544. https://doi.org/10.1007/s10212-022-00634-9

[82] Stephan Lewandowsky, Ullrich K.H. Ecker, and John Cook. 2017. Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era. *Journal of Applied Research in Memory and Cognition* 6, 4 (2017), 353–369. https://doi.org/10.1016/j.jarmac.2017.07.008

[83] Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest* 13, 3 (2012), 106–131. https://doi.org/10.1177/1529100612451018 arXiv:https://doi.org/10.1177/1529100612451018 PMID: 26173286.

[84] Stephan Lewandowsky and Sander van der Linden. 2021. Countering Misinformation and Fake News Through Inoculation and Prebunking. *European Review of Social Psychology* 32, 2 (2021), 348–384. https://doi.org/10.1080/10463283.2021.1876983 arXiv:https://doi.org/10.1080/10463283.2021.1876983

[85] Ming-Hui Li, Zhiqin Chen, and Li-Lin Rao. 2022. Emotion, analytic thinking and susceptibility to misinformation during the COVID-19 outbreak. *Computers in Human Behavior* 133 (2022), 107295. https://doi.org/10.1016/j.chb.2022.107295

[86] Q. Vera Liao and Wai-Tat Fu. 2013. Beyond the Filter Bubble: Interactive Effects of Perceived Threat and Topic Involvement on Selective Exposure to Information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2359–2368. https://doi.org/10.1145/2470654.2481326

[87] Q. Vera Liao, Wai-Tat Fu, and Sri Shilpa Mamidi. 2015. It Is All About Perspective: An Exploration of Mitigating Selective Exposure with Aspect Indicators. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1439–1448. https://doi.org/10.1145/2702123.2702570

[88] Falk Lieder, Thomas L Griffiths, and Ming Hsu. 2018. Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological review* 125, 1 (2018), 1.

[89] Scott O. Lilienfeld, Rachel Ammirati, and Kristin Landfield. 2009. Giving Debiasing Away: Can Psychological Research on Correcting Cognitive Errors Promote Human Welfare? *Perspectives on Psychological Science* 4, 4 (2009), 390–398. https://doi.org/10.1111/j.1745-6924.2009.01144.x arXiv:https://doi.org/10.1111/j.1745-6924.2009.01144.x PMID: 26158987.

[90] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. 2021. How WEIRD is CHI?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 143, 14 pages. https://doi.org/10.1145/3411764.3445488

[91] Jiqun Liu. 2023. Toward A Two-Sided Fairness Framework in Search and Recommendation. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval* (Austin, TX, USA) *(CHIIR '23)*. Association for Computing Machinery, New York, NY, USA, 236–246. https://doi.org/10.1145/3576840.3578332

[92] Elizabeth F. Loftus and John C. Palmer. 1974. Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior* 13, 5 (1974), 585–589. https://doi.org/10.1016/S0022-5371(74)80011-3

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

[94] Cameron Martel, Gordon Pennycook, and David G. Rand. 2020. Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications* 5, 1 (Oct. 2020), 47. https://doi.org/10.1186/s41235-020-00252-3

[95] Gillian Murphy, Emma Murray, and Doireann Gough. 2021. Attitudes towards feminism predict susceptibility to feminism-related fake news. *Applied Cognitive Psychology* 35, 5 (2021), 1182–1192. https://doi.org/10.1002/acp.3851 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.3851

[96] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175–220.

[97] Alexander Nussbaumer, Katrien Verbert, Eva-Catherine Hillemann, Michael A Bedek, and Dietrich Albert. 2016. A framework for cognitive bias detection and feedback in a visual analytics environment. In *2016 European Intelligence and Security Informatics Conference (EISIC)*. IEEE, 148–151.

[98] Aileen Oeberst and Roland Imhoff. 2023. Toward Parsimony in Bias Research: A Proposed Common Framework of Belief-Consistent Information Processing for a Set of Biases. *Perspectives on Psychological Science* 18, 6 (2023), 1464–1487. https://doi.org/10.1177/17456916221148147 arXiv:https://doi.org/10.1177/17456916221148147 PMID: 36930530.

[99] Margit E Oswald and Stefan Grosjean. 2004. Confirmation bias. *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory* 79 (2004), 83.

[100] Irene V Pasquetto, Briony Swire-Thompson, Michelle A Amazeen, Fabrício Benevenuto, Nadia M Brashier, Robert M Bond, Lia C Bozarth, Ceren Budak, Ullrich KH Ecker, Lisa K Fazio, et al. 2020. Tackling misinformation: What researchers could do with social media data. *The Harvard Kennedy School Misinformation Review* (2020).

[101] David Paternotte. 2015. Global times, global debates? Same-sex marriage worldwide. *Social Politics: International Studies in Gender, State & Society* 22, 4 (2015), 653–674.

[102] Gordon Pennycook, James Allan Cheyne, Nathaniel Barr, Derek J. Koehler, and Jonathan A. Fugelsang. 2015. On the reception and detection of pseudo-profound bullshit. *Judgment and Decision Making* 10, 6 (2015), 549–563. https://doi.org/10.1017/S1930297500006999

[103] Gordon Pennycook and David G. Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188 (2019), 39–50. https://doi.org/10.1016/j.cognition.2018.06.011 The Cognitive Science of Political Thought.

[104] Gordon Pennycook and David G. Rand. 2020. Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality* 88, 2 (2020), 185–200. https://doi.org/10.1111/jopy.12476

[105] Gordon Pennycook and David G. Rand. 2021. The Psychology of Fake News. *Trends in Cognitive Sciences* 25, 5 (May 2021), 388–402. https://doi.org/10.1016/j.tics.2021.02.007 Publisher: Elsevier.

[106] Suppanut Pothirattanachaikul, Takehiro Yamamoto, Yusuke Yamamoto, and Masatoshi Yoshikawa. 2019. 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* (Beijing, China) *(CIKM '19)*. Association for Computing Machinery, New York, NY, USA, 1653–1662. https://doi.org/10.1145/3357384.3357886

[107] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1650–1659.

[108] Tobias Richter. 2015. Validation and Comprehension of Text Information: Two Sides of the Same Coin. *Discourse Processes* 52, 5-6 (2015), 337–355. https://doi.org/10.1080/0163853X.2015.1025665

[109] Tobias Richter and Johanna Maier. 2017. Comprehension of Multiple Documents With Conflicting Information: A Two-Step Model of Validation. *Educational Psychologist* 52, 3 (2017), 148–166. https://doi.org/10.1080/00461520.2017.1322968

[110] Alisa Rieger. 2022. Interactive Interventions to Mitigate Cognitive Bias. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization* (Barcelona, Spain) *(UMAP '22)*. Association for Computing Machinery, New York, NY, USA, 316–320. https://doi.org/10.1145/3503252.3534362

[111] Alisa Rieger, Frank Bredius, Nava Tintarev, and Maria Soledad Pera. 2023. Searching for the Whole Truth: Harnessing the Power of Intellectual Humility to Boost Better Search on Debated Topics. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 248, 8 pages. https://doi.org/10.1145/3544549.3585693

[112] Alisa Rieger, Tim Draws, Mariët Theune, and Nava Tintarev. 2021. This Item Might Reinforce Your Opinion: Obfuscation and Labeling of Search Results to Mitigate Confirmation Bias. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media* (Virtual Event, USA) *(HT '21)*. Association for Computing Machinery, New York, NY, USA, 189–199. https://doi.org/10.1145/3465336.3475101

[113] Alisa Rieger, Tim Draws, Mariët Theune, and Nava Tintarev. 2023. Nudges to Mitigate Confirmation Bias during Web Search on Debated Topics: Support vs. Manipulation. *ACM Trans. Web* (nov 2023). https://doi.org/10.1145/3635034 Just Accepted.

[114] Alisa Rieger, Mariët Theune, and Nava Tintarev. 2020. Toward Natural Language Mitigation Strategies for Cognitive Biases in Recommender Systems. In *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*. Association for Computational Linguistics, Dublin, Ireland, 50–54. https://aclanthology.org/2020.nl4xai-1.11

[115] Emily Saltz, Claire R Leibowicz, and Claire Wardle. 2021. Encounters with Visual Misinformation and Labels Across Platforms: An Interview and Diary Study to Inform Ecosystem Approaches to Misinformation Interventions. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 340, 6 pages. https://doi.org/10.1145/3411763.3451807

[116] B. Schilit, N. Adams, and R. Want. 1994. Context-Aware Computing Applications. In *1994 First Workshop on Mobile Computing Systems and Applications*. 85–90. https://doi.org/10.1109/WMCSA.1994.16

[117] Norbert Schwarz, Eryn Newman, and William Leach. 2016. Making the Truth Stick & the Myths Fade: Lessons from Cognitive Psychology. *Behavioral Science & Policy* 2, 1 (2016), 85–95. https://doi.org/10.1177/237946151600200110

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

[119] Li Shi, Nilavra Bhattacharya, Anubrata Das, and Jacek Gwizdka. 2023. True or False? Cognitive Load When Reading COVID-19 News Headlines: An Eye-Tracking Study. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval* (Austin, TX, USA) *(CHIIR '23)*. Association for Computing Machinery, New York, NY, USA, 107–116. https://doi.org/10.1145/3576840.3578290

[120] Natalie J. Shook and Russell H. Fazio. 2009. Political ideology, exploration of novel stimuli, and attitude formation. *Journal of Experimental Social Psychology* 45, 4 (2009), 995–998. https://doi.org/10.1016/j.jesp.2009.04.003

[121] Herbert A Simon. 1957. A behavioral model of rational choice. *Models of man, social and rational: Mathematical essays on rational human behavior in a social setting* (1957), 241–260.

[122] Miroslav Sirota and Marie Juanchich. 2018. Effect of response format on cognitive reflection: Validating a two-and four-option multiple choice question version of the Cognitive Reflection Test. *Behavior research methods* 50 (2018), 2511–2522. https://doi.org/10.3758/s13428-018-1029-4

[123] Brendan Spillane, Séamus Lawless, and Vincent Wade. 2017. Perception of Bias: The Impact of User Characteristics, Website Design and Technical Features. In *Proceedings of the International Conference on Web Intelligence* (Leipzig, Germany) *(WI '17)*. Association for Computing Machinery, New York, NY, USA, 227–236. https://doi.org/10.1145/3106426.3106474

[124] N. Sriram and Anthony G. Greenwald. 2009. The Brief Implicit Association Test. *Experimental Psychology* 56 (2009), 283–294. Issue 4. https://doi.org/10.1027/1618-3169.56.4.283

[125] Namrata Srivastava, Rajiv Jain, Jennifer Healey, Zoya Bylinskii, and Tilman Dingler. 2021. Mitigating the Effects of Reading Interruptions by Providing Reviews and Previews. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 229, 6 pages. https://doi.org/10.1145/3411763.3451610

[126] C STANGOR and D MCMILLAN. 1992. MEMORY FOR EXPECTANCY-CONGRUENT AND EXPECTANCY-INCONGRUENT INFORMATION - A REVIEW OF THE SOCIAL AND SOCIAL DEVELOPMENTAL LITERATURES. *PSYCHOLOGICAL BULLETIN* 111, 1 (JAN 1992), 42–61. https://doi.org/10.1037/0033-2909.111.1.42

[127] Keith E Stanovich. 1999. *Who is Rational?: Studies of Individual Differences in Reasoning*. Psychology Press.

[128] K. E. Stanovich, R. F. West, and R. Hertwig. 2000. Individual Differences in Reasoning: Implications for the Rationality Debate? *Behavioral and Brain Sciences* 23, 5 (2000), 678–678.

[129] Natalie Jomini Stroud. 2008. Media Use and Political Predispositions: Revisiting the Concept of Selective Exposure. *Political Behavior* 30, 3 (2008), 341–366. http://www.jstor.org/stable/40213321

[130] Helge I. Strømsø, Ivar Bråten, and Tonje Stenseth. 2017. The role of students' prior topic beliefs in recall and evaluation of information from texts on socio-scientific issues. *Nordic Psychology* 69, 3 (2017), 127–142. https://doi.org/10.1080/19012276.2016.1198270 arXiv:https://doi.org/10.1080/19012276.2016.1198270

[131] Michael Sülflow, Svenja Schäfer, and Stephan Winter. 2019. Selective attention in the news feed: An eye-tracking study on the perception and selection of political news posts on Facebook. *new media & society* 21, 1 (2019), 168–190.

[132] Cass R. Sunstein and Richard H. Thaler. 2003. Libertarian Paternalism Is Not an Oxymoron. *The University of Chicago Law Review* 70, 4 (2003), 1159–1202. http://www.jstor.org/stable/1600573

[133] Charles S. Taber and Milton Lodge. 2006. Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science* 50, 3 (2006), 755–769. https://doi.org/10.1111/j.1540-5907.2006.00214.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-5907.2006.00214.x

[134] Yuko Tanaka, Miwa Inuzuka, Hiromi Arai, Yoichi Takahashi, Minao Kukita, and Kentaro Inui. 2023. Who Does Not Benefit from Fact-Checking Websites? A Psychological Characteristic Predicts the Selective Avoidance of Clicking Uncongenial Facts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 664, 17 pages. https://doi.org/10.1145/3544548.3580826

[135] Ben M Tappin, Gordon Pennycook, and David G Rand. 2020. Thinking clearly about causal inferences of politically motivated reasoning: why paradigmatic study designs often undermine causal inference. *Current Opinion in Behavioral Sciences* 34 (2020), 81–87. https://doi.org/10.1016/j.cobeha.2020.01.003 Political Ideologies.

[136] Predrag Teovanović, Goran Knežević, and Lazar Stankov. 2015. Individual differences in cognitive biases: Evidence against one-factor theory of rationality. *Intelligence* 50 (2015), 75–86. https://doi.org/10.1016/j.intell.2015.02.008

[137] R H Thaler and C R Sunstein. 2009. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Penguin Publishing Group.

[138] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology* 63, 1 (2012), 163–173. https://doi.org/10.1002/asi.21662 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.21662

[139] Cecilie Steenbuch Traberg and Sander van der Linden. 2022. Birds of a feather are persuaded together: Perceived source credibility mediates the effect of political

bias on misinformation susceptibility. *Personality and Individual Differences* 185 (2022), 111269. https://doi.org/10.1016/j.paid.2021.111269

[140] Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* 5, 2 (1973), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

[141] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131.

[142] Jan-Willem van Prooijen, André P. M. Krouwel, and Thomas V. Pollet. 2015. Political Extremism Predicts Belief in Conspiracy Theories. *Social Psychological and Personality Science* 6, 5 (2015), 570–578. https://doi.org/10.1177/1948550614567356 arXiv:https://doi.org/10.1177/1948550614567356

[143] Johan L.H. van Strien, Yvonne Kammerer, Saskia Brand-Gruwel, and Henny P.A. Boshuizen. 2016. How attitude strength biases information processing and evaluation on the web. *Computers in Human Behavior* 60 (2016), 245–252. https://doi.org/10.1016/j.chb.2016.02.057

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

[145] Joseph A. Vitriol, Joseph Sandor, Robert Vidigal, and Christina Farhart. 2023. On the Independent Roles of Cognitive & Political Sophistication: Variation Across Attitudinal Objects. *Applied Cognitive Psychology* 37, 2 (2023), 319–331. https://doi.org/10.1002/acp.4022 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.4022

[146] Ben Wang and Jiqun Liu. 2024. Cognitively Biased Users Interacting with Algorithmically Biased Results in Whole-Session Search on Debated Topics. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval* (Washington DC, USA) *(ICTIR '24)*. Association for Computing Machinery, New York, NY, USA, 227–237. https://doi.org/10.1145/3664190.3672520

[147] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300831

[148] P. C. Wason. 1960. On the Failure to Eliminate Hypotheses in a Conceptual Task. *Quarterly Journal of Experimental Psychology* 12, 3 (1960), 129–140. https://doi.org/10.1080/17470216008416717 arXiv:https://doi.org/10.1080/17470216008416717

[149] Magdalena Wischnewski, Rebecca Bernemann, Thao Ngo, and Nicole Krämer. 2021. Disagree? You Must Be a Bot! How Beliefs Shape Twitter Profile Perceptions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 160, 11 pages. https://doi.org/10.1145/3411764.3445109

[150] Haiping Zhao, Shaoxiong Fu, and Xiaoyu Chen. 2020. Promoting users' intention to share online health articles on social media: The role of confirmation bias. *Information Processing & Management* 57, 6 (2020), 102354. https://doi.org/10.1016/j.ipm.2020.102354

[151] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 436, 20 pages. https://doi.org/10.1145/3544548.3581318

[152] Qian Zhu, Leo Yu-Ho Lo, Meng Xia, Zixin Chen, and Xiaojuan Ma. 2022. Bias-Aware Design for Informed Decisions: Raising Awareness of Self-Selection Bias in User Ratings and Reviews. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 496 (nov 2022), 31 pages. https://doi.org/10.1145/3555597

5.3 Chapter Reflection

Computing systems and user interfaces can trigger the user's existing cognitive biases. However, how cognitive biases surface as a result of such triggering mechanisms is not straightforward. Cognitive biases only manifest in specific scenarios, which gives a fertile environment for these biases to resurface in the user-system interaction. In this chapter, we discuss the notion of bias susceptibility and its influence on the effects of confirmation bias when individuals consume information. Specifically, we explore factors that determine the user's bias susceptibility, which include the context of the user characteristics and the system interaction. User context describes individual differences in how users form their mental models when interacting with the systems. System-interaction context explains the relationship between the user and the system when they interact with each other. Addressing **(RQ 3)**, we conducted a user study that examined the influence of individual and contextual predictors on the effect of confirmation bias in three information consumption scenarios: seeking, recalling, and interpreting information. In this context, we deployed a subset of stimuli we used for the studies in Article II, comprising tweets expressing opinions on a polarising, divisive topic. We employed the ideological congruency between the user's beliefs and the content's stance as a proxy for confirmation bias. We measured its effects on the user's behaviour of information-seeking intention, information recall ability, and information interpretation rating.

The implication of this study is three-fold. First, we found that bias susceptibility predictors, such as the individual's thinking styles, political attitudes, topic interest, and the content's perceived issue strength, interacted with the effects of confirmation bias. Specifically, we found that an individual's tendency for low effortful thinking, the tendency for strong political attitudes (towards Liberal ideologies), and the content's strong perceived issue strength all amplified the effects of confirmation bias. Second, designers can consider the bias susceptibility factors when building and deploying interventions to mitigate the undesired effects of confirmation bias. Platforms can soften the issue strength of the content as well as avoid curating content items that carry a risk of amplifying the user's confirmation bias. Platforms leverage the user's individual differences, identify user profiles that exhibit more susceptibility to cognitive biases, and personalise interventions, safeguarding mechanisms, or educational prompts that cater to specific users. The awareness of bias susceptibility will allow designers to build *context-aware* interventions on media platforms and help fortify users from manipulation. Third, we remark that the choice of task design and modality can influence the effect of confirmation bias. We found that (the same group of) individuals showed different bias susceptibility across different tasks. The nature of the task or interaction scenario can introduce cognitive biases (i.e., systematic effects) as a confounding variable. These additional biases or effects can skew the user behaviour and, therefore, the strength of the effect of confirmation bias across different behavioural measurements we deployed in our study.

To conclude this chapter, we point out the influence of user- and context-related factors on the effects of cognitive biases, rendering users more or less susceptible to exhibiting biases. The role of bias susceptibility not only informs the design of interventions to mitigate cognitive biases but also allows us to detect the occurrences of cognitive biases more precisely. The findings in this chapter complement and extend the implications of Chapter 4, which suggests that cognitive biases surface in a fertile environment and are expressed in detectable patterns of physiological signals, e.g., fNIRS. This chapter explores what factors render conditions fertile for cognitive biases. Chapter 3 suggests that HCI researchers focus on understanding cognitive biases in the interaction to generate considerations for the design of computing systems. This chapter thereby enriches the understanding of bias susceptibility and casts guidelines for designing computing systems that take cognitive biases into account. In the next chapter, we discuss how we can leverage the design into bias-aware systems that guide users away from problematic behaviour when navigating the online world.

Chapter 6

Towards Designing Bias-Aware Technologies

6.1 Introduction

When people consume online media, cognitive biases can be problematic as individuals offload their cognitive processing and give up critical thinking. Misinformation often contains deceptive messages that tap into people's cognitive biases. Specifically, these messages cling to ideologically polarising, divisive issues like climate change, abortion rights, or vaccine effectiveness. Thus, individuals resort to many of their existing heuristics, including the ideological congruence between their beliefs and the content's stance. In this realm, confirmation bias drives users to seek and engage predominantly with content that conforms to their beliefs. In Chapters 4 and 5, we study this heuristic and discuss methods and factors to help capture the effects of cognitive biases in this interaction context. Furthermore, Chapter 5 discuss considerations for designing interventions to mitigate the adverse effects of confirmation bias by taking bias susceptibility into account.

The goal of this thesis is to design computing systems that address and adapt to cognitive biases in users. The HCI community has increasingly formed an understanding of how cognitive biases manifest in HCI and how designs can better keep up with the dynamics of these biases. In Chapter 3, our scoping review of cognitive bias studies in HCI outlines three connecting narratives: HCI researchers (1) *quantify* the effects of cognitive biases to (2) better *understand* their effects in human-computer interaction and subsequently (3) form *designs of computing systems* that address these biases in the real world. In Chapter 4 and 5, we address the angles of quantifying and understanding the effects of cognitive biases. In this chapter, we extend beyond the contributions presented in the previous chapters by tackling the angle of designing systems to address cognitive biases in the real world.

Addressing **(RQ 4)**, we present discussions in three academic workshops (at CHI 2020 [35], CHI 2021 [36], and Dagstuhl seminar 2022 [37]). These workshops brought together experts from HCI, cognitive psychology, behavioural science, and law to explore the role of cognitive biases in human-computer interaction. The workshop discussion was centred around the topic of online misinformation, as it is one of the concerns introduced by cognitive biases. We summarise the workshop discussions and propose a research agenda guiding the design of computing systems that equip people with skills and affordances to make informed decisions when navigating the online world. Also, the workshop discussions signal a need to use interdisciplinary knowledge and theory to guide the design of computing systems that address cognitive biases in their users. Ultimately, we pave the way towards the design of bias-aware technologies, which address and take into account cognitive biases that exist in users and emerge during their interaction with computers. We describe in detail workshop discussions, their outcomes, and implications in Article IV.

6.2 Article IV

Nattapat Boonprakong, Benjamin Tag, and Tilman Dingler. 2023. Designing Technologies to Support Critical Thinking in an Age of Misinformation. In *IEEE Pervasive Computing*, July-Sept 2023. IEEE, Vol. 22, No. 3, pp. 8-17. https://doi.org/10.1109/MPRV.2023.3275514

THEME ARTICLE: AUGMENTED COGNITION

Designing Technologies to Support Critical Thinking in an Age of Misinformation

Nattapat Boonprakong, *University of Melbourne, Melbourne, VIC, 3010, Australia*
Benjamin Tag, *Monash University, Clayton, VIC, 3800, Australia*
Tilman Dingler, *University of Melbourne, Melbourne, VIC, 3010, Australia*

Algorithms increasingly curate the information we see online, prioritizing attention and engagement. By catering to personal preferences, they confirm existing opinions and reinforce cognitive biases. When it comes to polarizing topics such as climate change or abortion rights, the combination of algorithmic information curation and cognitive biases can easily skew our perception and, thus, undermine our critical thinking abilities while creating a thriving ground for misinformation. To curb the spread of misinformation, a research agenda is needed around the interplay between cognitive biases, computing systems, and online platform design. In this article, we synthesize insights from a workshop series, propose a research agenda, and sketch out a blueprint for technologies to support critical thinking through the lens of human–computer interaction and design. We discuss the affordances of online media and how they could prioritize teaching users how to spot misinformation better and conduct themselves in online environments.

Pervasive technology has enabled us to access information at any time. Facing vast amounts of information each day, people often adopt mental shortcuts or heuristics to filter and sift through content more effectively, a phenomenon often described as *cognitive biases*.[1] Coined by Kahneman and Tversky, cognitive biases involve the use of personal preferences and prior experiences to simplify the complexity of information processing. At the same time, they can hinder our critical thinking and, therefore, our ability to make informed decisions. Prominent examples include *confirmation bias* (predominantly seeking out information that confirms existing views), *cognitive dissonance* (repudiation of information that does not fit into preconceived notions), and *the continued influence effect* (previously learned misinformation continues to influence an individual's beliefs and attitudes even after correction).

By optimizing recommendation systems around the content's "stickiness" and keeping users engaged at all costs, computing systems—systems that run on algorithms—tend to reinforce cognitive biases since the algorithms largely cater to users' interests and beliefs.[2] When it comes to contentious topics like climate change, abortion rights, or immigration policies, this becomes a serious problem as computing systems and cognitive biases work together to skew our perception and foster the creation of so-called filter bubbles and echo chambers,[3] where like-minded people discuss topics from one congruent angle and mutually confirm each others' standpoints. Ultimately, these silos of comfort run the risk of barring us from multifaceted discussions and contribute to alienation by creating an "us" versus "them" narrative.

The manifestation and reinforcement of biases also exacerbate the spread of misinformation. Defined as false or misleading information, misinformation prompts people to doubt the nature of truth, give up on analysis, and process information by using heuristics instead.[4] Confirmation bias, for example, helps lubricate the spread of misinformation as it prompts people to not only seek predominantly one-sided information, but also to accept information without necessarily considering its veracity. Other cognitive biases like the continued influence effect also make misinformation difficult to debunk as misinformation tends to linger around even after it has been corrected.

1536-1268 © 2023 IEEE
Digital Object Identifier 10.1109/MPRV.2023.3275514
Date of publication 1 June 2023; date of current version 6 October 2023.

The interplay between humans and computing systems, however, allows researchers to intervene on two fronts. On the side of computing and pervasive systems, researchers have developed solutions to counter the spread of misinformation. Pattern recognition and natural language processing (NLP) yield automatic approaches to detect misinformation and fake news.[5] In recent years, we have seen fact-checking services being introduced on social media platforms like Facebook and Twitter. Yet, these approaches are largely based on ground truth data provided by human labellers. This has led to a head-to-head race between human fact-checkers and the speed of misinformation production. Additionally, some online content, e.g., satire, is challenging to spot and correctly handle, sometimes even for humans. Effectively tackling the spread of misinformation through algorithmic solutions alone is, therefore, not always feasible.[6]

On the human side, researchers have been focusing on understanding how people evaluate and interact with information online,[7] and on designing interventions to nudge and empower people to make deliberate decisions in the digital world.[8] Such research has been mainly driven by practitioners and researchers from different disciplines including computer science, behavioral economics, cognitive psychology, and political science. The problem of misinformation can only be addressed in a truly interdisciplinary fashion as it involves the technology through which fake news propagates, the human who creates, receives, and shares it, and the regulatory bodies who are looking for ways of reeling in its spread.[9] By focusing on the user-side perspective, equipping people with critical thinking abilities would help fortify them against falling victim to misinformation.

With the field of human–computer interaction (HCI) at the forefront of the design and development of user-facing computing systems, we bear special responsibility for working on solutions to mitigate problems arising from misinformation and bias-enforcing interfaces. At the same time, the interdisciplinary nature of HCI provides a fruitful climate to involve experts from different areas to collectively discuss, create, and assess countermeasures against the spread of misinformation from a well-rounded perspective.

In this article, we discuss a research agenda of technologies to combat the spread of misinformation arising from a series of three workshops we organized over recent years. Each workshop brought together researchers from various disciplines to identify and rethink the mechanisms of pervasive computing systems behind the spread of misinformation. We summarize and present the various challenges and options for the development and implementation of technology solutions that tackle the spread of misinformation and help foster critical thinking abilities in people. As an outcome of our workshop series, we propose *affordances of online media* that help equip people with skills to critically engage with information and conduct themselves in the digital world.

BACKGROUND

The research community has developed and evaluated a wide range of approaches and tools to identify and tackle misinformation. Nevertheless, the question remains why people are susceptible to misinformation and why countermeasures are not more effective. Therefore, we set out to not only investigate computing systems and their role but also include insights derived from cognitive science that enable us to better understand the role of human cognition.

As one of the pioneering researchers of counteracting misinformation, Lewandowsky et al.[9] pointed out that a special feature of misinformation is its presentation as a conspiracy theory. By providing simple solutions to complex problems, conspiracy theories are easier to grasp than often convoluted explanations from official sources. Additionally, nonexperts often fall back to simplifications due to a lack of knowledge and expertise in a rather complex domain, such as climate change.[7] A recent countermeasure introduced by Lorenz-Spreen et al.[8] uses nudging and boosting techniques to support online users in making informed decisions, for example, showing the number of likes of an article against the total number of readers. The idea behind nudges, i.e., subtle stimuli that trigger specific behaviors, comes from behavioral economics research.[10] The effectiveness of nudges has further been studied by Pennycook et al.[7] who show that such nudges, e.g., giving an accuracy reminder when reading online news, can support online users in making decisions on what information to share and what not.

Research in HCI has tackled the problem mainly by designing novel user interfaces. Caraban et al.[11] provided insights on how to effectively design nudges used in digital environments by targeting cognitive biases in users. In their follow-up work,[12] the authors narrow in on nudging interventions in the context of misinformation with the aim of helping people to discern credibility and balance their information diet. Other countermeasures have been proposed by Jeon et al.[13] and targeted psychological inoculation and gamification techniques to fortify online users against misinformation.

As stakeholders of the misinformation crisis, social media platforms have begun to put HCI research into practice. Twitter, for example, introduced its misinformation policy to slow down the spread of misleading content: the platform equips labels on the content items and nudges users toward trustworthy information sources.[a]

While the HCI community has proposed a few approaches for interventions aimed at fostering positive behavior change, prior works are largely lacking a common research target: which problems should be prioritized? The algorithms or the human? The spreader or the recipient? The incentive structures of social media platforms or informing regulation?

With its user-centered design approach and ability to rapidly prototype and test interventions, the HCI community is uniquely positioned to create and study countermeasures against misinformation. Such interventions, however, should be grounded in supporting theory from the fields of psychology, policy-making, and behavioral sciences.

In this article, we lay out the discussions we had with experts from different disciplines about the role of cognitive biases in the spread and reception of misinformation. Over the course of three workshops, we explored the reciprocal relationship between computing systems and users' cognitive biases with the goal of formalizing a research agenda to investigate that relationship in more detail and identify measures to effectively curb the spread of misinformation.

WORKSHOPS

To explore the role of cognitive biases in the interactions with computing systems, we brought together experts from different disciplines over the course of three workshops in recent years. We initially started with a workshop hosted at the ACM Conference on Human Factors in Computing (CHI) to broadly scope the relationship between cognitive biases in people and computing systems. A second CHI workshop subsequently leads us to narrow in on the challenge of misinformation and the role of biases as catalysts for its spread. Realizing that these issues reach well beyond the scope of HCI research, we organized a three-day Dagstuhl seminar, which opened up the discussion to a broader audience, including disciplines such as psychology, behavioral economics, and policy-making.

The entirety of this workshop series served the common goal of creating a research agenda around cognitive biases and their role in interacting with pervasive computing systems as they relate to the spread of misinformation.

Workshop 1: CHI 2020

The first workshop in our series focused on the idea that cognitive biases are omnipresent when we interact with computing systems. The goal was to scope which types of biases are of interest in the context of user-system interaction and how computing systems may contribute to their creation or reinforcement. Therefore, there is a need to clearly define the bias, when it occurs, and how it could be detected. We specifically focused on the research around cognitive biases through the lens of HCI. The workshop entitled *Detection and Design for Cognitive Biases in People and Computing System* was held at CHI in 2020.[14]

The workshop themes were kept explicitly broad as we wanted to understand the scope of the role of cognitive biases in interacting with computing systems. Hence, the themes focused on the understanding, utilization, detection, and mitigation of cognitive biases in people and computing systems. We invited workshop participants from the HCI community to submit a position paper on their ideas, 12 of which we subsequently discussed in the one-day workshop with 19 participants.

The position paper contributions[b] were mainly concerned with the mechanisms, applications, and mitigation strategies of cognitive bias in different areas: media studies, behavioral economics, education, healthcare, and computer science. Several papers echoed the previously mentioned bias amplification by system algorithms and interfaces. One paper pointed out a case study of Youtube recommendation algorithms, which tend to expose users to cater predominantly videos align with content users have already watched and those from authoritative sources, e.g., major news outlets. Thus, the author discussed that the algorithm risks amplifying the users' confirmation bias and anchoring bias, respectively.

By observing the nature of biases, some papers proposed solutions to avoid or mitigate negative effects from cognitive biases, for example, altering the interface design, providing alternative perspectives, or creating bias-reflection tools. One paper discussed the continued influence effect in the use case of

[a][Online]. Available: https://help.twitter.com/en/resources/addressing-misleading-info

[b]Position papers (CHI 2020) are accessible from http://critical-media.org/cobi/schedule.html

mobile learning. Some papers also discussed computational methods to detect cognitive biases. By leveraging sensing techniques, such as eye tracking, body movement, or physiological measures, we can detect discrepancies within the human–machine interaction (e.g., information browsing behavior or healthcare patient–provider interaction). Subsequently, we can "sense" the occurrences of biases in people and design interventions that give constructive and timely feedback about the bias that users may not be aware of.

In addition, we included two action group activities (5-6 people in each group), where participants received a quest problem and brainstormed solutions in a group. We used a combination of *Miro* and *Zoom* to allow participants to ideate and collaborate with each other.

The first action group activity focused on identifying research questions around cognitive biases. We asked participants to list 1) which problems can they solve, 2) which problems are in need of a solution, and 3) which problems need to be defined first. Participant groups prioritized research around the detection and mitigation of cognitive biases and formulated the following research questions:

- **How can we detect biases? How can we estimate the impact of biases?** Cognitive biases often occur without us being consciously aware of them. While it is crucial to mitigate the negative effects of biases, we first have to understand when biases occur, what triggers them, and how we could *quantify* their effects. Quantification of biases is a challenging task as reliable ground truth is hard to collect. Mechanisms for reliably inducing biases would be required to systematically study their occurrences.
- **Which biases are most promising to start with?** Addressing biases that have the most impact would best help alleviate the overall negative effects of cognitive biases. Moreover, these biases are more tractable for self-awareness as they occur rather often. The position papers to this workshop highlighted some potential candidates of cognitive biases to start with. Confirmation bias and anchoring bias, for example, are prominent in the settings of online news and media consumption: confirmation bias makes people susceptible to an echo chamber effect, where they expose themselves to predominantly one-sided information; anchoring bias, meanwhile, impacts how people believe misinformation as they tend to rely on the first piece of information regardless of its veracity.
- **How do we help people to become aware of their own biases? Would reflecting on biases help mitigate their effects?** Making people aware of their biases and supporting their metacognition could help reduce the negative effects arising from biases and fortify people from making fallacious decisions. This would lead us either back to the challenge of detecting biases or we would need to develop a framework in which people actively reflect on their decision-making and on the factors that contribute to it in order to identify and understand the role that cognitive biases may play in these decisions.

Subsequently, we conducted another action group activity to identify how to measure and mitigate cognitive biases. Each participant chose a cognitive bias and listed theories or technologies to address and mitigate them. We found that most proposed mitigation strategies were based around *nudges*.

Research in HCI has long employed nudges as a way to induce behavior change in people by changing their environment or choice architecture.[11] Nudging leverages the knowledge of our cognitive biases and supports people in making better decisions.[10] Here are some of the nudging techniques that were discussed in the workshop:

- **Friction**: Introducing friction in computing system interactions aims at slowing down the user's decision-making. This slow-down intends to make users rethink their available choices and question their initial gut reactions. Friction could hence reduce the effects of cognitive biases by allowing users to engage in thoughtful reflection and introspection and subsequently make more deliberate decisions. An example of introducing friction would be to limit the number of likes on Twitter per time interval or enable commenting on a tweet after a certain time has elapsed.
- **Education tools**: Learning is often affected by our innate biases. *The continued influence effect* says that the first information we consume tends to stick even when disproven later on.[15] This could be utilized for inoculation strategies, where the first encounter of a concept is used to bolster users' critical thinking abilities. Roozenbeek and Van der Linden[16] proposed the use of games to playfully inoculate people against fake news and build their critical thinking abilities.

Overall, we observed a broad range of research interests as cognitive biases occur not only during

various information consumption tasks but also across different application areas, such as information systems, finance, healthcare, and education. The research space for cognitive biases is therefore not only understudied but also vast. To serve our original research goal to mitigate the spread of misinformation, the research scope needed to be significantly narrowed down.

Workshop 2: CHI 2021

The realization from the previous workshop that one of the most effective ways to protect people from misinformation is to bolster their critical thinking abilities lead to the second workshop in our series entitled *Technologies to Support Critical Thinking in an Age of Misinformation*,[17] which we held at CHI 2021. We narrowed the scope of the role cognitive biases play in the interaction with computing systems to the reception and spread of misinformation. Similarly, we invited workshop participants across the HCI community to lay down their arguments through position paper submissions concerned with topics related to system and user interface design that address issues of misinformation.

Similar to the outcomes from the first workshop, there have been attempts to mitigate cognitive biases or leverage them for the better good through promoting prosocial behaviors—social behaviors that benefit other people or society as a whole. Examples include mitigating the spread of fake news, depolarizing people, and cutting down toxicity on the Internet.

Throughout the workshop, we provided 16 participants with a framework to discuss interventions based on *the Prosocial Design Network platform.*[c] The platform had been codeveloped by one of the workshop coorganizers and aims to gather the collective effort of academics and designers across disciplines who work on technology solutions to help promote prosocial behavior: for example, being more critical and less toxic. The network's website hosts a categorized collection of tested and untested technology interventions based on existing scientific evidence.

The motivation behind the prosocial design network is that most existing solutions rely on common sense or intuition about what will work. A wide range of tools and approaches to curb the spread of fake news exist but they do not always work. Some interventions could risk causing a *backfire effect*—exacerbating existing biases. For example, in case a person has already been convinced by misinformation, confronting them with opposing arguments often results in the solidification of the existing beliefs. Designing technologies to combat misinformation is therefore not trivial and needs to be well-grounded in psychological science.

Workshop activities were structured in the same way the Prosocial Design Network challenges were and centered around technology interventions and interface design. The one-day workshop was kicked off with a keynote talk by Anastasia Kozyreva on *Confronting Digital Challenges with Cognitive Tools*.[18] The talk summarized three categories of technology interventions: technocognition (interventions inspired by cognitive science[9]), nudging (interventions that change people's behavior in a predictable way through choice architecture[10]), and boosting (interventions that foster people's cognitive and motivational competencies[19]). The talk was followed by the workshop participants' submission paper presentation and an action group activity (3-4 people in each group) on designing concrete interventions.

We received four position paper submissions[d] that were concerned with the interplay between computing systems and cognitive biases in people. The first two papers sought to understand the characteristics of the interaction between users and misinformation: triggering negative emotions and cognitive biases. One paper investigated people's behavioral and emotional expressions regarding fake news from publicly available social media data. The findings suggested that fake news tends to generate negative emotions in users rather than exert positive influence. Similarly, another study evaluated the design elements of the *Parler* social media platform. Their findings suggest that the choice of design elements can make the platform conducive to the flow of misinformation. For example, the prioritization of images over text content can amplify emotional responses in users; and the limited search functionality hinders the user's ability to seek information from alternate perspectives.

Two papers proposed interventions to counter the spread of misinformation. One paper was concerned with emancipatory design aiming to support media literacy on social media platforms. By exposing users to increased transparency, we can mitigate the negative effect of cognitive biases and help them nurture critical thinking abilities. For example, by showing users the existence of their filter bubbles or their political partisanship, we can invite users to reflect on their

[c][Online]. Available: https://www.prosocialdesign.org

[d]Position papers (CHI 2021) are accessible from http://critical-media.org/chi21/schedule.html

otherwise hidden biases and intervene, e.g., by suggesting exposure to alternative viewpoints. Another paper discussed three prophylactic interventions to counter fake news on social media: inoculation, fostering media literacy, and imposing transaction cost economics. These interventions induce cognitive effort in message evaluation and thus reduce the likelihood of interacting with social media content with low credibility.

Overall, the position paper submissions imply that technology is a double-edged sword: interface designs and algorithms can be weaponized to amplify users' cognitive biases and exacerbate the spread of fake news. In contrast, the same technologies can be used to nurture critical thinking abilities in people and provide safeguards against misinformation.

The task for the action group participants was to design an intervention that supports critical thinking in people. Each intervention had to be described along the two dimensions *target* and *method*. The target of the intervention included building media literacy, breaking the echo chamber, reducing fake news, or countering hate speech; the methods by which this is accomplished included prebunking, providing friction, providing labels/extra information, promoting social interaction, providing media transparency, and modifying news feeds. Participants described their interventions in the spirit of the Prosocial Design library by naming "what it does" and "how it works."

Here are some examples of interventions created by the workshop participants:

- **Echo chamber warning** (Break the echo chamber × Labeling): A web browser widget that gives feedback about the diversity of the user's news exposure, calculated by their previous browsing history. The widget tracks the stance of websites visited by the user and acts as a nudge for a more balanced information diet.
- **Insist on comprehension before sharing** (Media literacy × Providing friction): Before users can post a comment on an online article, they must pass a quick comprehension quiz on what they have read.
- **Reputation coloring** (Media literacy × Providing media transparency): The user community ranks each message/post/tweet according to its veracity. At the same time, each user is accompanied by badges and subtle color coding indicating the frequency of their use of accurate, false, or misleading information.
- **Systematically inoculate people against misinformation** (Reduce fake news × Prebunking): Mandatory briefings through videos about what the characteristics of bad information are for new users. This would provide users with "immunity" against believing and propagating misinformation.

A common theme we observed in the proposed interventions was adding delay in the interaction. Friction, as a nudging technique, aims at helping users employ more deliberate decision-making. To combat the spread of misinformation, friction has already been deployed on many social media platforms. Twitter, for example, created explicit friction by prompting users to quote a tweet with commentary instead of retweeting during the 2020 US presidential election. Facebook imposed a message forwarding limit on its instant messaging platforms—WhatsApp and Facebook Messenger.

Some discussions held during the workshop suggested that computer science researchers, software developers, and designers alone struggle to provide a comprehensive solution to the problem of misinformation. It became clear that regulators and industry partners are playing crucial roles in this fight against misinformation as well. The outcomes from this workshop suggested the need to open up this ongoing discussion to a broader group of actors, especially those outside of the computer science community.

Workshop 3: Dagstuhl Seminar 2022

We ran the third workshop of our series as a Dagstuhl seminar.[20] To comprehensively discuss solutions to the problems of misinformation from a well-rounded perspective, we brought together researchers and practitioners from a wide range of areas, such as computer science, behavioral science, media studies, psychology, as well as people from industry and lawmaking.

Throughout the three-day seminar, we organized a number of talks and group activities that addressed misinformation in different aspects. On the first day, participants jointly collected and discussed a range of challenges that misinformation on digital platforms presents. Throughout the day, we decided on three final themes around which we formed the following three working groups: 1) regulation, 2) critical thinking, and 3) human factors & platforms. Each group consisted of around five individuals.

As a group, participants first discussed challenges in mitigating the spread of misinformation. Subsequently, we employed the 5-whys activity to let them search for the root cause of the problem they

identified earlier. Starting from the challenge statement (e.g., "people interpret information in their own ways"), participants repeatedly asked a "why" question to discover the fundamental or hidden assumption of the problem.

Arising from the activities of day one, each group laid out the challenges as follows:

- **The regulation group** pointed out that problems and difficulties in creating effective regulation are the competence, the evidence, and the lack of consensus between different stakeholders. Specifically, law and policymakers are unable to continuously catch up with the ongoing technological changes.
- **The critical thinking group** suggested that we not only need to nurture critical thinking and media literacy in people but also encourage people to step out of their silos of comfort and make an effort to get to know and understand the perspectives of other individuals.
- **The human factors & platforms group** proposed that there are information and presentation factors, as well as the crowd and platform governance mechanisms, which cause different individuals to interpret information in their own ways.

On the second day of the workshop, we let participants generate new ideas by using the reverse brainstorming protocol. Instead of thinking about direct solutions to a problem, the protocol is to brainstorm ways to create the problem, and then reverse them into new and alternative solutions. On the following day, we arranged the solution exploration activity to let participants further transform their ideas from earlier activities into more concrete research questions and executable research projects.

In sum, we highlight a number of research proposals spanning from the working groups[e] as follows:

- **A framework to write better policies for technological regulation**: We need a way to inform policy-makers of effective regulatory measures. At the same time, we need to also make sure that regulations do not silence unwanted voices and break the fundamentals of a pluralistic society. To do so, we need an eco-system that facilitates reproduction studies and foster collaboration and knowledge exchange between regulators, academia, and industry.
- **Tools and methodologies to benchmark and evaluate interventions**: Researchers and designers have proposed digital interventions to combat the spread of misinformation and foster critical thinking (e.g., the previously mentioned Prosocial Design Network). Yet, there is a lack of direction and methodology to make sure that these solutions are effective. We, therefore, need standardized ways to test the effectiveness of new interventions to allow benchmarking and comparison. Knowing the effectiveness and impact of different interventions would not only help inform researchers and designers but also people in law-making who enforce regulations and public measures.
- **Teaching people skills to thrive through misinformation**: To fortify people from falling victim to false information, we need to equip people with critical thinking and media literacy skills. Critical thinking should not be only taught in schools but also on social media platforms as flagging falsified, misleading, or harmful content could act as an *opportunistic education* and, more or less, impact the users' critical thinking skills. Through group discussions, we collated a list of techniques and interventions to teach critical thinking abilities as follows:
  1) *Innoculation:* Build psychological immunity against misinformation. By getting exposed to common mechanisms of creating and spreading misinformation, receivers of inoculation training acquire the ability to spot and resist manipulation.
  2) *Pre/Debunking:* We can correct misinformation after exposure (debunking) or before exposure (prebunking). Yet, misinformation correction should be done carefully as it bears the risk of backfire effects.
  3) *Building Empathy:* Seeing the world from someone else's perspective is crucial to understanding where the other is coming from. Connecting and empathizing are the key to meaningful conversations, especially around critical topics.
  4) *Platform Moderation:* To facilitate healthy discourses on a contentious topic, moderation might be necessary to bring people with different views to the table.
- **Integrating critical thinking and media literacy into the education system** While media literacy training has made its way into middle school

[e]For the full report of activities and results, see the Dagstuhl report 22172.[20]

education, the quickly changing nature of the online discourse requires a frequent revisiting of these contents and techniques.

- **Improving academia-industry collaboration**: Researchers in academia and industry are required to collaborate and better identify mutual needs and rights. More importantly, the access to information collected by social media platforms like Facebook is deemed extremely helpful to researchers, allowing them to create insights into human behavior.
- **Understanding people through objective measures**: There has been limited understanding of why false information is taken as fact and disseminated by humans. The availability of sensors and pervasive devices, such as smartphones, nowadays offer researchers to make sense of human cognition. Physiological markers, such as electrodermal activity and heart rate, provide objective and nonintrusive measures of the human cognitive state and thus may allow researchers to investigate how misinformation is being perceived mentally.

RESEARCH AGENDA

Arising from our workshop series, we derived the following two themes for a research agenda: to equip people with critical thinking and prosocial skills and to understand and design affordances of the online environment.

Equipping People With Better Online Skills

How can we support people not falling victim to misinformation? The three workshops echoed that, to effectively combat the spread of misinformation, we need to equip people with better online skills, i.e., how to behave as healthy, critical, and prosocial online citizens. To do so, we need technologies that teach and nurture these skills in people. We have seen collective efforts like the Prosocial Design Network in which technology interventions are systematically tracked and classified by an interdisciplinary group of researchers.

Our workshop series outlined the need to establish a systematic design space for systems and interfaces that equip users with these skills. In recent years, different technology interventions have been proposed and deployed on platforms. However, existing solutions tend to be more based on common sense than grounded in theory and empirical evidence. The voices from our workshop participants suggested three considerations for designing technology interventions.

1) **What skills do we need to equip users with?** What skills do people need to conduct themselves safely in today's online spaces? Our workshops outlined some examples of the necessary skills: critical thinking and prosocial abilities. When designing an intervention, the first thing one should consider is its concrete goal (for example, reducing the spread of misinformation, pushing people out of their silos of comfort, or mitigating online toxicity). Knowing what skills or goals are important helps designers and researchers identify interventions that should be performed.
2) **How to ensure the intervention is effective?** While we have seen interventions being employed on platforms, it is yet unclear whether they worked in practice. There is a need to establish a systematic research design to (objectively) measure the effectiveness of the proposed interventions. We suggest that one solution can be to define a set of metrics with respect to the goal of interventions. Yet, future research needs to investigate how well these metrics measure effectiveness.
3) **Is the intervention backed by scientific theories?** An effective intervention needs supporting grounds in psychological science. Kozyreva et al.[18] summarized three countermeasures supported by behavioral and cognitive psychology: nudging, boosting, and technocognition. With interventions being informed by theory, we can be more confident that they are less likely to backfire.

Creating Affordances of Online Media

The second theme revolves around the use of design to point users toward more critical discourse with information. Our workshop participants pointed out that computing systems and algorithms played a crucial role in amplifying people's biases, meanwhile, we could also use the same technologies to mitigate the negative consequences arising from them. By focusing on the system-side perspective, we could create so-called *affordances* of online media, making users more likely to critically evaluate information and behave prosocially.

The concept of affordances was originally coined by James Gibson in 1966 as the possible actions that users perceive when interacting with an object. We envision affordances of online media to prompt users

to critically reflect and avoid bad online behaviors, such as sharing fake news. Interface design elements should hence shift users' behaviors in a benign way with the goal of reducing misinformation and creating more informed online discussions. While nudges and affordances of online media are similar, they are different in the sense that nudges alter the choice architecture, while affordances serve as a broader approach to fostering positive behaviors by revealing the choice architecture.

What are the affordances of online media that could foster these positive behaviors? The goal would be to teach critical thinking skills and "good citizenship" behaviors in situ, i.e., as users interact with computer systems and navigate the information space. Affordances, hence, nudge users toward learning and incorporating these behaviors as opposed to extended inoculation techniques, where users would go through explicit schooling. An example would be to point out fallacies in online information, which signal users not only to avoid sharing less credible information but also to learn about news characteristics to help them to recognize misinformation, even when explicit labels are removed.

Equipping people with critical thinking and prosocial skills through interacting with computing systems is in line with the concept of augmented cognition, where systems and designs build and boost users' capabilities to technology skills and drive better debates in the online world. While algorithms and cognitive biases still tend to sort us into silos of comfort and administer one-sided information diets, the technology solutions discussed in our workshops present some concrete interventions targeted at helping people overcome their biases and sharpen the critical thinking skills that protect them and their decision-making from misinformation.

CONCLUSION

To explore the relationship between cognitive biases and the interaction with computing systems, we brought together researchers and practitioners from different disciplines to identify and discuss countermeasures against misinformation. The research agenda focuses on the detection of the occurrence of biases and how to mitigate their effects. By prioritizing critical thinking, interventions can be directly integrated into the design of computing systems and teach users how to defend themselves against misinformation by building innate skills, i.e., augmenting their cognitive capabilities. These design affordances of online media work in situ, i.e., nudge users toward better online behaviors and navigating the online space more safely. Specifically, we arrive at two crucial research questions: What are the affordances of online media that could teach people to become more critical and prosocial in the online world and how could we provide these affordances?

The space of technology solutions lies on a continuum of user agency. At one end of the spectrum, nudging interventions subtly alter the users' behaviors but tend to take away the user's agency as the approach itself may exploit their unconscious biases. On the flip side, pedagogical interventions—for example, giving formal education on media literacy or inoculating people against misinformation—give higher user agency and have been shown to be highly effective but come with the caveat of requiring users' willingness and time to be educated.

Affordances, on the other hand, can directly be incorporated into the design of online media to educate or nudge users by offering possible actions. They can also be mandated by regulatory bodies. They basically relieve users from formal media literacy training while finding a compromise between their effectiveness and users' willingness to be educated.

To effectively counter the problem of misinformation, we need coordinated efforts from not only academic researchers but also regulators and policymakers. With a framework of how to design and evaluate interventions along with a mechanism to inform policy-makers, we may be able to bridge the knowledge gap between academia and regulation. Ultimately, interventions can be designed, tested, and deployed to a wide audience, thus equipping the public with the necessary skills to thrive in the digital world.

ACKNOWLEDGMENTS

The author would like to thank A. Vargo, K. Kise, A. Dengel, P. Lorenz-Spreen, E. Karapanos, S. Knight, S. Lewandowsky, and L. Devillers for coorganizing and driving the mentioned workshops. They also thank all workshop participants for their contributions and feedback and would like to acknowledge the Dagstuhl Seminar 22172 where this publication was conceived.

REFERENCES

1. A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," *Science*, vol. 185, no. 4157, pp. 1124–1131, 1974.
2. E. Hussein, P. Juneja, and T. Mitra, "Measuring misinformation in video search platforms: An audit study on youtube," *Proc. ACM Hum.-Comput. Interact.*, vol. 4, no. CSCW1, May 2020, doi: 10.1145/3392854.

3. E. Pariser, *The Filter Bubble: What the Internet is Hiding from You*. London, U.K.: Penguin U.K., 2011.
4. C. Lampe, "Beyond bots and buttons–new directions in information literacy for students," *IEEE Pervasive Comput.*, vol. 20, no. 4, pp. 79–81, Oct.–Dec. 2021.
5. K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *SIGKDD Explor. Newsl.*, vol. 19, no. 1, pp. 22–36, Sep. 2017, doi: 10.1145/3137597.3137600.
6. E. Strickland, "Ai-human partnerships tackle "fake news": Machine learning can get you only so far-then human judgment is required - [news]," *IEEE Spectr.*, vol. 55, no. 9, pp. 12–13, Sep. 2018.
7. G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, and D. G. Rand, "Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention," *Psychol. Sci.*, vol. 31, no. 7, pp. 770–780, 2020.
8. P. Lorenz-Spreen, S. Lewandowsky, C. R. Sunstein, and R. Hertwig, "How behavioural sciences can promote truth, autonomy and democratic discourse online," *Nature Hum. Behav.*, vol. 4, no. 11, pp. 1102–1109, 2020, doi: 10.1038/s41562-020-0889-7.
9. S. Lewandowsky, U. K. Ecker, and J. Cook, "Beyond misinformation: Understanding and coping with the "post-truth" era," *J. Appl. Res. Memory Cogn.*, vol. 6, no. 4, pp. 353–369, 2017.
10. R. H. Thaler and C. R. Sunstein, *Nudge: Improving Decisions About Health, Wealth, and Happiness* (Series A New York Times Bestseller). London, U.K.: Penguin Publishing Group, 2009.
11. A. Caraban, E. Karapanos, D. Gonççalves, and P. Campos, "23 ways to nudge: A review of technology-mediated nudging in human-computer interaction," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, 2019, pp. 1–15, doi 10.1145/3290605.3300733.
12. L. Konstantinou, A. Caraban, and E. Karapanos, "Combating misinformation through nudging," in *Proc. 17th IFIP TC 13 Int. Conf. Hum.-Comput. Interact.*, 2019, pp. 630–634.
13. Y. Jeon, B. Kim, A. Xiong, D. LEE, and K. Han, "Chamberbreaker: Mitigating the echo chamber effect and supporting information hygiene through a gamified inoculation system," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW2, pp. 1–26, Oct. 2021, doi: 10.1145/3479859.
14. T. Dingler, B. Tag, E. Karapanos, K. Kise, and A. Dengel, "Workshop on detection and design for cognitive biases in people and computing systems," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, 2020, pp. 1–6, doi: 10.1145/3334480.3375159.
15. S. Lewandowsky, U. K. H. Ecker, C. M. Seifert, N. Schwarz, and J. Cook, "Misinformation and its correction: Continued influence and successful debiasing," *Psychol. Sci. Public Int.*, vol. 13, no. 3, pp. 106–131, 2012, doi: 10.1177/1529100612451018.
16. J. Roozenbeek and S. Van Der Linden, "The fake news game: Actively inoculating against the risk of misinformation," *J. Risk Res.*, vol. 22, no. 5, pp. 570–580, 2019.
17. T. Dingler, B. Tag, P. Lorenz-Spreen, A. W. Vargo, S. Knight, and S. Lewandowsky, "Workshop on technologies to support critical thinking in an age of misinformation," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, 2021, pp. 1–5, doi: 10.1145/3411763.3441350.
18. A. Kozyreva, S. Lewandowsky, and R. Hertwig, "Citizens versus the internet: Confronting digital challenges with cognitive tools," *Psychol. Sci. Public Int.*, vol. 21, no. 3, pp. 103–156, 2020, doi: 10.1177/1529100620946707.
19. R. Hertwig and T. Grüne-Yanoff, "Nudging and boosting: Steering or empowering good decisions," *Perspectives Psychol. Sci.*, vol. 12, no. 6, pp. 973–986, 2017, doi: 10.1177/1745691617702496.
20. T. Dingler, B. Tag, and A. Vargo, "Technologies to support critical thinking in an age of misinformation (Dagstuhl seminar 22172)," *Dagstuhl Rep.*, vol. 12, no. 4, pp. 72–95, 2022. [Online]. Available: https://drops.dagstuhl.de/opus/volltexte/2022/17281

NATTAPAT BOONPRAKONG is currently working toward the Ph.D. degree with the School of Computing and Information Systems, University of Melbourne, VIC 3010, Australia. His research interest involves empathic computing and cognition-aware systems, in particular the quantification of cognitive biases. He is the corresponding author of this article. Contact him at nboonprakong@student.unimelb.edu.au.

BENJAMIN TAG is a lecturer in the Immersive Analytics Lab, Monash University, Clayton, VIC 3800, Australia. His research interests include human-AI interaction, digital emotion regulation, and human cognition with a special focus on inferring mental state changes from data collected in the wild. Contact him at benjamin.tag@monash.edu.

TILMAN DINGLER is a computer scientist and senior lecturer with the University of Melbourne, VIC 3010, Australia. His research focuses on cognition-aware systems, the detection of cognitive states, and the adaptation of computing systems to aid their users. Contact him at tilman.dingler@unimelb.edu.au.

6.3 Chapter Reflection

Computing systems can pose problematic effects in the real world. Especially, the spread of misinformation can be attributed to the cognitive biases individuals use to fast-track their information processing and economise their cognitive resources. Cognitive biases are even more pronounced as media platforms could trigger and amplify their effects. For example, platforms employ recommendation algorithms to optimise user engagement. Such algorithms learn the user preferences and beliefs, and, therefore, predominantly cater to content items that strengthen the user's confirmation bias. We discussed in the previous chapters that the design of computing systems can introduce devastating effects – either by mistake of the designer or deliberate abuse. In this chapter, we ask whether the very same solution can be used to address problems it has contributed.

We address **(RQ 4)** through a qualitative coding analysis of three workshop discussions that brought together interdisciplinary experts to draft a blueprint of technologies to address cognitive biases and their contribution to the spread of misinformation. We derived two themes for a research agenda. First, we can design systems that equip people with skills to navigate the online world and avoid engaging in problematic behaviours, for example, sharing unverified information. Interventions to train such skills span from simple nudging prompts, psychological inoculation, or systematic education modules. We point out open questions about what skills individuals should be trained in and, thus, be embedded as part of the intervention design. Moreover, how can we define and measure the effectiveness of interventions? Both questions will inform the design of interventions that are goal-oriented and effective. Secondly, we can create technologies with affordances for users to be critical, unbiased. While computing systems can trigger and amplify cognitive biases, the same solutions can be reverse-engineered to mitigate biases and keep people making rational decisions. More specifically, the design of computing systems can rather nudge users away from problematic behaviour. We ask a key question: What are the affordances of online media that could foster non-problematic, positive behaviour? Additionally, we take a special note that the proposed research agenda cannot proceed without the coordinated effort of multi-disciplinary researchers. Designing systems to address real-world concerns related to cognitive biases needs not only HCI knowledge but also insights from behavioural science, cognitive psychology, as well as ethics.

The findings of Article IV lay the ground for future research into the design of bias-aware systems. Addressing, mitigating, and leveraging cognitive biases is a challenge. It involves understanding a nuance of the human mind with respect to the complexity of the real world. The proposed research agenda resonates with the findings in the previous chapters. For example, chapter 3 suggests that, to advance the understanding of cognitive biases in the interaction, the field of HCI should better engage with behavioural science and psychology. Chapters 4 and 5 provide cues for the design of bias-aware systems, being the indicators for the occurrences of cognitive biases and content-related factors that amplify the effects of cognitive biases. In the next chapter, we will convene the findings from our original research contributions and reflect on how they address our research questions.

This page is intentionally left blank.

Chapter 7

Discussion and Future Directions

Humans are hardwired to employ mental shortcuts to effectively process information, make decisions, and navigate the real world. Cognitive biases emerge as these shortcuts systematically influence human decision-making and behaviour. When humans interact with computers, cognitive biases can surface in the interaction as users employ mental shortcuts to navigate the user interface while optimising their cognitive resources. There is a growing discussion among HCI researchers about how cognitive biases influence the interaction between users and systems. This thesis, therefore, sets out to chart a systematic understanding of the effects of cognitive biases in HCI. We present the original research contributions of this thesis in Chapters 3, 4, 5, and 6. In this chapter, we discuss our findings in relation to the literature around cognitive biases, address the research questions proposed in Chapter 1, and set out avenues for future research.

7.1 Cognitive Biases in Human-Computer Interaction

Cognitive biases influence how we behave, navigate, and make decisions in the real world. When humans come into contact with computers, these biases also affect how we interact with computing systems. While the role of the rational-bounded human mind clearly contributes to the manifestation of cognitive biases, we lack a sufficient understanding of how the design of computing systems and user interfaces impacts cognitive biases that surface in the interaction between humans and computing systems. The field of HCI is at the forefront of designing interactive systems – HCI researchers advance the understanding of human factors revolving around the human use of computing systems, especially how these systems can impact human factors. To address the research gap, we ask the first research question:

RQ 1: How are cognitive biases studied in HCI?

7.1.1 Essential Components to Study Cognitive Biases

In Chapter 3, we build upon a scoping review of cognitive bias studies in HCI (Article I) and outline the narratives of how HCI researchers engage with cognitive biases. The findings suggest that (A) Computing systems can trigger and amplify existing cognitive biases in users; (B) Designers take advantage of cognitive biases and build systems that trigger biases and subsequently steer user behaviour; (C) HCI researchers develop tools and methods to quantify and observe the effects of cognitive biases to derive designs that take biases into account. We leverage these narratives into a summary of how HCI researchers study cognitive biases. Revisiting **(RQ 1)**, we propose that HCI researchers study cognitive biases not only to understand them as

a human factor but also to inform the design of computing systems that take cognitive biases into account. HCI researchers develop tools and methods to capture the effects of cognitive biases in human-computer interaction, allowing them to form a clearer understanding of the impact of cognitive biases in the interaction. The understanding of cognitive biases, therefore, allows HCI researchers to design computing systems that adapt to cognitive biases that can arise during the interaction. We visualise this summary in Figure 1.1, which allows us to identify three essential elements to study cognitive biases in HCI, which pose as the research gap we aim to address in Chapters 4, 5, and 6:

- **Tools & Methods:** We lack tools and methods for HCI researchers to quantify, study, and observe the effects of cognitive biases.
- **Understanding:** We need more research investigating how computing systems can trigger existing cognitive biases in users.
- **Design:** We need to formulate considerations for designing bias-aware computing systems, which address, leverage, and take into account cognitive biases in users

7.1.2 The Definition of Cognitive Biases in HCI

In Chapter 3, we discuss the divisive status of the definition of cognitive biases in HCI. In psychology, researchers are divided over the rationality war [149] between two schools of thought: Tversky and Kahneman's *Heuristics and Biases* [178] and Gigerenzer's *Fast and Frugal Heuristics* [52]. The former suggests cognitive biases as systematic errors in judgment, while the latter believes these biases are useful strategies that help humans to make optimal decisions in the real world. Taken together, researchers may take cognitive biases as the human mind's undesirable weakness or useful strength – a double-edged sword.

Do cognitive biases mostly imply *undesirable* effects in HCI? Chapter 3 points out that the majority of articles in the corpus (39 papers) seek to mitigate the adverse effects of cognitive biases, while only 21 papers focus on utilising the effects of cognitive biases. The two focuses of study reflect different sentiments HCI researchers have regarding the notion of cognitive biases as either something to be *corrected* or to *leveraged*. We draw a parallel to the rationality war debate, which divides how psychologists view cognitive biases. If cognitive biases were errors, they should be corrected. If these biases were useful, they should be leveraged. In a recent work, Baigelenov et al. [10] found that most publications in visualisation research put more emphasis on cognitive biases as flaws of human decision-making. Their analysis of sentiment towards cognitive biases showed that, among 112 papers analysed, only 3 papers adopted a balanced perspective, and no paper took a positive stance. We argue that the term "cognitive bias" can hint at an emphasis on human judgment errors, which Kahneman and Tversky first proposed in their seminal work [178]. Indeed, Pohl [136] surveyed the function of cognitive biases and accounted that most psychologists viewed cognitive biases as "built-in errors of the human information-processing systems." However, the review, which was published in 2004, may have reflected an outdated, conventional view of psychologists towards the notion of cognitive biases.

Can cognitive biases be *beneficial* in HCI? Cognitive biases are inherent to humans and, therefore, cannot be completely separated. Chapter 3 shows that two ways cognitive biases can benefit HCI. First of all, we found that HCI researchers utilise cognitive biases to guide user behaviour. This is in line with Thaler and Sunstein's nudge theory [169, 175]. Because cognitive biases influence decision-making, designers can tap into the user's cognitive biases and steer their decisions in a predictable way. Secondly, the investigation of cognitive biases in the interaction helps us shape the design of computing systems that better align with the user's cognitive biases. Designers can identify cognitive biases and minimise those effects that are harmful or confounding (i.e., introducing undesired side effects). Specifically, we can diagnose elements of computing

systems (e.g., user interfaces and algorithms) that trigger cognitive biases to better understand the mechanism behind how cognitive biases are triggered, amplified, and mediated.

Recent research in psychology moves towards a more balanced emphasis on the notion of cognitive biases as hardwired, adaptive features of the human mind [70, 131]. These biases produce systematic, reproducible patterns of real-world human behaviour [11]. The field of HCI has emerged to study human factors in computing systems, with a goal of improving the design and its user experience. In Chapter 3, we reflect on the practice of HCI researchers in deriving design considerations from studying cognitive biases in the interaction. Taking a neutral stance that cognitive biases are systematic effects hardwired to the human mind, we can define cognitive biases as a methodological lens to explain user behaviour and a guide to designing computing systems that take these biases into consideration. More specifically, designers can achieve their intended outcomes of the system interaction and avoid unexpected, confounding, or detrimental effects triggered by computing systems.

Can we use cognitive biases as a *theory* in HCI? Scholars in HCI adopt theories, concepts, and constructs from behavioural science and psychology as a method to guide their research [145]. Our scoping review indicates that the notion of cognitive biases is not widely employed in HCI studies. Our article inclusion criteria restrict articles discussing specific forms of cognitive biases but lacking reference to the notion of cognitive biases, e.g., design fixation [188], conformity [198], overreliance [99, 200], priming [108, 165], framing [29, 67], or placebo effect [182]. Some related concepts are also discussed in HCI research, for example, decision-making fairness [60, 201], bounded rationality [97, 115, 129], human reasoning [33], nudging [13], unconscious human behaviour [163], and algorithmic conformity [203]. While the exclusion of such articles is a limitation of our scoping review, it signals that cognitive biases span and explain a diverse number of effects and phenomena in HCI – some of which are not acknowledged as cognitive biases. We share a similar sentiment with Kliegr et al. [101], who voiced that there is a plethora of behavioural phenomena not accounted for as cognitive biases. Eronen and Bringmann [42] and Hagger [65] argued that the field of psychology redundantly invents definitions, constructs and theories. In the same manner, we argue that HCI scholars engage with a large space of related, overlapping psychological phenomena – they quantify, understand, and build designs of computing systems that take into account such effects. We envision that our proposed summary of how HCI researchers study cognitive biases in Figure 1.1 can be extended to explain how HCI researchers engage with other psychological phenomena.

7.1.3 Tools and Methods to Quantify the Effects of Cognitive Biases

Cognitive biases occur pervasively in human-computer interaction. Capturing the effects of cognitive biases will allow HCI researchers to precisely and closely study their impact on the user interaction with systems and develop a clearer understanding of their effects. However, cognitive biases generally happen without our awareness. Asking users whether they have cognitive biases is not a feasible way to gauge the presence of cognitive biases. On the other hand, we can observe their effects through behavioural and physiological expressions. Because cognitive biases are part of our mental process for effective information processing (i.e., using mental shortcuts), these biases reflect on our physiological responses and, subsequently, behaviour. However, the research gap indicates that we do not have a clear understanding of the indicators for the occurrences of cognitive biases. Therefore, we ask the second research question:

RQ 2: What are the indicators for the occurrences of cognitive biases when they manifest in HCI?

In Chapter 4, we conducted two user studies to explore the indicators for the occurrences of cognitive biases when people encounter different opinions on a divisive, polarising topic. Individuals employ the congruence between the content's ideological stance and their existing beliefs as their heuristic. Addressing

(RQ 2), our findings highlight that physiological expressions, especially hemodynamic responses, showed significant changes when our participants read dissenting opinions and, therefore, yielded as reliable indicators of the occurrences of cognitive biases. We also found a trend in skin conductance level changes when our participants read dissenting opinions, suggesting that cognitive biases could also influence EDA signals.

To the best of our knowledge, this is the first study to apply physiological sensing in the form of fNIRS and EDA to measure the effects of cognitive biases in human-computer interaction. We build on prior work on the effects of cognitive dissonance [18, 96, 195], which suggested that individuals have higher neural activations when they face information that challenges their beliefs. In addition, our findings augment the potential of EDA in capturing the dissonance arousal, which is a concept in psychology that dissenting information causes an increase in arousal [202]. Echoing prior work by Ploger et al. [135], we also detected an insignificant trend in EDA-derived skin conductance levels. Through a pilot study of 7 participants, Ji et al. [82] demonstrated that when comprehending information on a difficult topic, electroencephalography (EEG) showed a clearly higher neural activation, and the skin conductance level (SCL), which can be derived from EDA, showed greater variability and fluctuation. Moreover, they reported that SCL exhibited fewer individual differences as compared to EEG. We envision future research to further investigate the potential of physiological signals for quantifying the effects of cognitive biases in HCI.

Our findings offer tools and methods to quantify the effects of cognitive biases in HCI. Specifically, our studies offer a two-step guideline for conducting cognitive bias studies in HCI. First, we induce cognitive biases using stimuli that elicit the use of heuristics when interacting with computing systems. Second, we probe the effect of cognitive biases using behavioural or physiological measurement tools. In Chapter 4, we induced cognitive biases through exposing individuals to reading opinions on a divisive, polarising topic (e.g., climate change, feminism, or abortion rights). In this regard, we focused on the congruence between the participants' existing beliefs and the content's ideological stance as a heuristic for comprehending the bias-inducing stimuli. We then employed an eye-tracking camera, fNIRS, and EDA to probe the effects of cognitive biases when our participants read the stimuli. Future research can employ this guideline to improve their internal validity by making sure that cognitive biases happen in the interaction before conducting further investigation, like mitigating or utilising their effects.

7.1.4 Understanding Cognitive Bias Susceptibility

In Chapter 4, while we found significant effects of ideological congruency on hemodynamic activity, they are more pronounced in individuals who exhibited low interest in the topic corresponding to the bias-inducing stimuli. Prior research also suggested that individual characteristics, such as reflective thinking tendency or need for cognition, can influence the effects of cognitive biases [21, 141] and the individual's receptivity to bias mitigation interventions [101, 119]. Specifically for physiological activity, Ji et al. [81, 82] emphasised that individual differences can confound the observable effects of cognitive biases. Therefore, we speculate that the user and interaction contexts play a crucial role in influencing the effects of cognitive biases. In Chapter 3, we point out a lack of studies investigating cognitive biases in conjunction with user- and context-related factors, suggesting that we lack sufficient understanding of how cognitive biases operate across different users and interaction contexts. To address this research gap, we ask the third research question:

RQ 3: What and how factors for user- and interaction-context influence the occurrences of cognitive biases when they manifest in HCI?

In Chapter 5, we introduce the concept of cognitive bias susceptibility, suggesting that cognitive biases do not always surface in the interaction as they are subject to the influence of the user and interaction contexts. Addressing **(RQ 3)**, we conducted a user study to examine the interaction effects between confirmation bias

and predictors for cognitive bias susceptibility and found that both individual factors (tendency for effortful thinking and political concordance) and contextual factors (the user's topic interest, the content's perceived issue strength, and the task design) are factors for cognitive bias susceptibility, influencing the effects of confirmation bias. The understanding of cognitive bias susceptibility informs the design of computing systems to better adapt to the user's cognitive biases. In Chapter 5, we discuss such design considerations in relation to media platform designs where interventions can be catered specifically to user and context characteristics that are fertile for cognitive biases, for example, users with a strong political inclination, or information conveying a strong issue.

To the best of our knowledge, our study is the first to focus on the interplay between the effects of cognitive biases and factors for cognitive bias susceptibility. We augment the discussion in the HCI community around the need to consider the context where cognitive biases manifest. The works of Rieger et al. [140, 142, 143] and Graells-Garrido et al. [61] envisioned that interventions to mitigate cognitive biases should shift away from one-size-fits-all approaches. Liu [112] recommended the need to consider user- and system-side factors when examining fairness in human decision-making, as these factors influence how humans make decisions. In explainable AI research, Kliegr et al. [101] emphasised the need to consider context when studying cognitive biases, as the same piece of information can trigger different cognitive biases in different contexts. Chapter 3 voices that multiple cognitive biases can oppose or reinforce each other. In Chapter 5, we argue that interventions to mitigate the effects of cognitive biases should be context-aware. Specifically, designers should consider user- and context-related factors when curating and deploying interventions on end-users.

Our contributions enrich the understanding of how cognitive biases operate in human-computer interaction. It not only informs the design considerations for computing systems regarding cognitive biases but also helps complement findings in behavioural science and psychology. In Chapter 3, we echo Hekler et al. [72], who voiced that the fields of HCI and behavioural science can benefit from each other. The field of HCI can help verify concepts in behavioural science through rapid prototyping and user testing. Ultimately, HCI research can elevate theoretical concepts into practical implications.

7.1.5 Designing Bias-Aware Computing Systems

The contribution of Chapter 3 also extends into how computing systems should be designed to adapt to the human mind and cognition. Recent research has proposed the notion of *bias-awareness* [113, 204], which refers to the ability to detect, understand, and take into account cognitive biases. In Chapter 4, we propose that HCI researchers can build computing systems that are bias-aware. With the ability to identify what and when cognitive biases manifest in the interaction, we can apply interventions to precisely and effectively address their effects. Chapter 5 suggests that the manifestation of cognitive biases is subject to individual and contextual factors. As well, these factors can affect a user's receptiveness to bias mitigation interventions [34, 119]. We propose that computing systems that address the effects of cognitive biases should adapt to the contexts of user and system interaction. Yet, we do not have sufficient understanding of how we can equip computing systems with bias-awareness. Hence, we ask the fourth research question:

RQ 4: What are the considerations for the design of computing systems that take cognitive biases into account?

In Chapter 6, through a qualitative analysis of three workshop discussions concerning the issue of cognitive biases in HCI, we propose two recommendations for designing interventions to address cognitive biases that promote the spread of misinformation. First of all, designers can provide affordances in the design of online media to shift the user behaviour towards informed, reflective thinking. We derive Gibson's concept of affordances [50], suggesting that an environment offers possible actions that users perceive. In the same vein,

the design of online media can invite users to engage in reflective thinking rather than relying on intuition. For example, the design of media platforms may assist users in consuming media without provoking their cognitive biases. In Chapter 3, we discuss interventions to mitigate the effects of cognitive biases as cognitive assistance to shift users away from mental shortcuts that produce harmful effects. In the realm of misinformation, we can design online media to support the user's cognitive processing and decision-making when they navigate the online world. Secondly, we propose that equipping people with skills to navigate the online world will help address harmful cognitive biases. These interventions will train users to be more resilient against misinformation and online manipulation. This is in line with prior research in HCI [106, 184, 197] suggesting that pedagogical tools can teach users critical thinking to avoid cognitive biases. More importantly, we suggest that interventions should have clear objectives (i.e., what skills to equip users with?), metrics for effectiveness (i.e., how do we ensure interventions are effective?), and well-defined theoretical grounds.

To address **(RQ 4)**, we can generalise the affordances of online media into the affordances of computing systems. Specifically, we can design computing systems as an environment that invites users to make informed decisions, as well as shifts them away from intuition and automatic thinking. In Chapter 3, we found that the majority of cognitive biases identified can be attributed to *Too Much Information* and *The Need To Act Fast* causes, implying computing systems tend to overwhelm users with information and force them to make quick decisions. Therefore, the affordances of the present-day computing systems may not allow users to engage with systems critically and deliberately. Because users have limited cognitive resources and time, they can only afford to employ mental shortcuts when navigating the user interface and, therefore, exhibit cognitive biases. On the other hand, we can design computing systems that allow users – who have limited cognitive, memory, time, and knowledge resources – to navigate without overrelying on mental shortcuts. Computing systems can shift away from imposing cognitive demands to augmenting the user's cognitive, memory, time, and knowledge capacities. Ultimately, we can *bridge the gap* between the design of computing systems and the design of the human mind.

Page [131] suggested that the design of the human mind no longer keeps up with the environment we live in today. As part of the evolution, we developed mental shortcuts as optimal strategies to navigate the real world. However, the environment we originally adapted to has experienced a rapid change. Cognitive biases surface because the environment we live in is not a Stone Age cave, but rather, we are in the world of pervasive computing devices that are not designed to accommodate our existing cognitive biases and align with our cognitive processes. We argue that the gap can be minimised if we equip computing systems with bias-awareness. In particular, the design of computing systems can provide affordances for users to make decisions without spending too much of their own cognitive resources.

7.2 Societal and Ethical Considerations of Bias-Aware Systems

Our findings further suggest that HCI researchers should study cognitive biases with care. Earlier in Chapter 3, we discussed cognitive biases in HCI as a double-edged sword. We claim that HCI researchers can leverage the understanding of cognitive biases to develop bias-aware systems that better adapt to the user's cognitive processes. The exploitation of cognitive biases, however, can offer both benefits and harm. Designers can build computing systems that avoid triggering undesired effects, act as cognitive assistants, and induce positive behaviour change. On the other hand, cognitive biases can be deliberately exploited to manipulate user behaviour and decision-making without their awareness, such as sludges [174], dark patterns [117], and social engineering [19]. With this thesis exploring the role of cognitive biases in HCI, we need to rethink the ethics and impact of bias-aware systems on individuals and society.

7.2.1 Societal Considerations

Computer scientists and engineers tend to *optimise* computing systems to better serve humans; they have a technical perspective of building a system that is faster, more efficient, and more accurate [193]. Selbst et al. [153] argued that such designs fail to account for the fairness of sociotechnical systems, thus introducing problematic effects on individuals and society. Similarly, by overlooking cognitive biases that can surface, elements of computing systems can introduce systematic side effects. For example, while deep learning systems can provide a highly accurate prediction, users may not trust its output because of the lack of transparency of black-box models [147]; conversational AIs are trained to be a human partner, but they tend to conform with the user's beliefs [203]; recommendation systems are optimised to cater predominantly to the user's preferences, but they amplify the user's confirmation bias [152, 154] or the order effect [64].

We argue that many instances of cognitive biases in the real world originate from the incentive structure of today's computing systems. For example, social media platforms deploy recommendation algorithms to keep users engaged [75]. These algorithms learn from the user's latent preferences and beliefs and subsequently curate suitable content items. As a result, the design intended to optimise user engagement can amplify the user's confirmation bias [15]. Similarly, online dating services are designed to help people with distant social backgrounds to form relationships. However, these services tend to reinforce people's racial prejudices and stereotypes [114], and therefore, e.g., foster homophily and divide people [47]. In another example, instant messaging services are designed to enhance mutual awareness and relational closeness [85]. While these services are asynchronous in nature, they tend to introduce information overload and the sense of urgency to respond, leading to delays in reply [109] and harming interpersonal relationships [28].

While evidence suggests that these cognitive biases introduce harmful effects, should the solution to reduce such biases be a revision of the incentive structure itself? We propose that there is a trade-off between optimising the intended functionality of a system and minimising the effects of undesired cognitive biases. For example, removing read receipts on instant messaging may lose its purpose for real-time communication, but help alleviate effects like the urgency to reply. Dropping the accuracy of system recommendations may reduce user satisfaction, though it may weaken the effects of confirmation bias. We believe there is a need to reconsider the incentives of system design with regard to human cognitive limitations and bounded rationality. If a system aims to maximise its intended objectives, does it demand our cognitive processing excessively so that we offload it using mental shortcuts? We envision that bias-awareness adds a dimension to designing computing systems under the abovementioned trade-off. Perhaps we can find a compromise (e.g., an intervention) that keeps the system functionality while allowing less harmful biases to surface.

7.2.2 Ethical Considerations

There are ethical concerns for bias-aware computing systems, which we voiced in Chapters 3, 4, 5, and 6. First, harnessing cognitive biases can imply manipulation of people's behaviour. For computing systems that trigger cognitive biases with intention, e.g., nudging and dark patterns, users are generally unaware that their biases are being utilised for an intervention. Psychologists criticised nudging as a form of *benevolent paternalism* [84, 87, 170] – an authority to restrict an individual's choices of action for the best interest of the people – and its lack of transparency [17, 66, 105]. Bovens [17] suggested that nudges are more effective if they work in the dark, without the user's awareness. Hansen and Jespersen [66] defined nudges according to their mode of thinking engaged and transparency, suggesting that nudges span from manipulative to reflective interventions. In the realm of HCI, Caraban et al. [22] pointed out that nudges can be intrusive and run the risk of harming user autonomy. The visibility and the ease of opting out of nudges will determine the ethics of nudges.

Similar to nudges, some dark patterns tap into cognitive biases to manipulate user behaviour [117]. The

deceptive, opaque, and manipulative nature of dark patterns has been long criticised as an ethical issue for design [62]. Caragay et al. [24] proposed that dark patterns are deviations from the normative of ethical software design, which violate the user's expectation of systems. We argue that dark patterns operate by triggering harmful cognitive biases that render users vulnerable to manipulation. While technologies triggering cognitive biases in people can offer benefits, they should be carefully designed in a way that does not violate user expectations, intrude on user autonomy, and trigger detrimental effects.

Secondly, the understanding of cognitive biases can be abused to steer individuals' behaviour. One can design a system that triggers cognitive biases to systematically manipulate the individual's ideological and behavioural tendencies. We take a classic example of the Nazi German propaganda in thr 1930s and 40s where citizens were exposed to predominantly one-sided news. As a result, the majority of Germans who grew up under the regime had a skewed belief towards anti-semitism [185]. A similar technique has been employed in the success of worldwide propaganda, instilling a belief tendency in a mass population. In a recent account, Burda et al. [19] suggested that social engineering applications target people's cognitive biases and manipulate their behaviour, for example, phishing and pretexting scams. The deliberate exploitation of cognitive biases, as seen in historical propaganda and modern social engineering, highlights the ethical imperative to foster critical thinking and media literacy to safeguard individuals from manipulative influences.

7.3 Future Directions

Building upon the findings of this thesis, we discuss avenues for future research that are worth investigating. First, we highlight threats to the ecological validity of cognitive bias studies in HCI and potential ways to address them. Secondly, we call attention to aligning the cognition gap between the design of computing systems and the human mind.

7.3.1 Improving Ecological Validity

Cognitive biases are mental strategies we use to navigate the real world. In contrast, limited research in HCI observes cognitive biases in a real-world setting. Indeed, most studies tend to investigate the effects of cognitive biases in a controlled environment. As in Chapter 4, we discussed that we can study cognitive biases in HCI by inducing their effects while isolating away *other* confounding effects. In Chapter 5, acknowledging that individual and contextual factors can influence the effects of cognitive biases, we take such factors into account as cognitive bias susceptibility. While this thesis outlines a research methodology for studying cognitive biases, there is a plethora of threats to the generalisability of the findings. For example, the experimenter effect [133] (i.e., the presence of an experimenter influences the participant's behaviour), the amplitude of the cognitive biases (i.e., the effects of cognitive biases are too small to be observed), and, more importantly, the presence of multiple cognitive biases. In Chapter 3, we voiced that different cognitive biases can surface at the same time and interact (i.e., cancel or reinforce) with each other. For example, reading an ideologically dissenting opinion may cause both confirmation bias and disconfirmation bias [40] (the tendency to scrutinise statements dissenting from one's beliefs). It is, therefore, a challenge to clearly separate individual cognitive biases from each other as they result from an individual's subjective judgment.

We recommend future research to rethink how we can study the effects of cognitive biases. While cognitive biases are originally coined as systematic patterns of deviation from the norm of rationality, these biases lead to both optimal decisions and flawed judgments. Bertrand et al. [14] suggested the need to label biases as *beneficial*, *detrimental*, or *neutral*. In a cognitive bias experiment, the user's observable response is a *net* effect after taking into account the effects of different cognitive biases that reinforce and cancel each other. Similarly, in Chapter 4, we used a blanket term "cognitive biases" to cover multiple cognitive biases (beyond

confirmation bias, which commonly happens in information consumption) that may arise when individuals read and evaluate opinions. We considered the overall effect of multiple cognitive biases on the behavioural and physiological responses. In sum, we envision that researchers can extend cognitive bias studies into more realistic settings by moving beyond the focus on one specific cognitive bias.

7.3.2 Bridging the Cognition Gap between Humans and Computers

The notion of cognitive biases exists as a theory to explain that humans often deviate from the norm of rational judgment. In this thesis, we learned that these biases can be triggered by computing systems and pose problematic effects in the real world. However, recent discourses in psychology have discovered their limits. Page [131, 132] suggested that, rather than focusing on how different cognitive biases manifest, we should investigate why these biases occur. It is unreasonable that, despite human decision-making abilities having evolved over natural selection, humans fundamentally make flawed judgments. Depending on the norm, context, individual, and environment, the same cognitive bias can produce optimal decisions or abnormalities (i.e., introducing problematic effects). Especially, Page advocated that the research community needs to focus on the big picture of the mismatch between the design of our mind and the environment of the present day.

Similarly, research in HCI has voiced the need to bridge the cognition gap between humans and computers [14, 171], as the digital environment we live in does not match with our cognitive processes. In Chapter 3, we documented the mention of cognitive biases and categorised them according to Benson's problem attribution of cognitive biases [11], which explained how mental shortcuts help address each of the fundamental cognitive demands. Our findings, therefore, provide evidence that computing systems can impose information and memory overload, time constraints, and ambiguity in decision-making. Especially, most of the identified cognitive biases can be attributed back to the two former cognitive demands. In the present-day complex environment, such excessive demands can hinder the alignment between the human mind and computers.

While bias-aware computing systems can help designers better align computing systems and users' existing cognitive biases, we can generalise this challenge to be matching the design of computing systems and the design of the *adaptive* human rationality. We echo Page's advocacy [131] for the need to shift away from a specific focus on cognitive biases in one HCI application scenario. In Chapter 3, in the same vein, we call for future research to consider cognitive biases beyond a specific context, generalising findings and design considerations in a context-agnostic manner. In Chapter 5, we voiced that the application context can influence the effects of cognitive biases. Cognitive bias susceptibility plays a crucial role in hindering or amplifying the effects of cognitive biases, explaining that these biases do not always manifest. We suggest that cognitive bias studies in HCI may not replicate because of the differences in study settings and participant demographics. In the same way, psychology and behavioural science have experienced a replication crisis [88, 118].

We believe there is a need to further investigate elements of computing systems that not only trigger cognitive biases but also introduce users to *excessive* cognitive demands. Specifically, we need to rethink the system designs to address human challenges in real-world decision-making. The concept of providing affordances in bias-aware systems, which we introduced in Chapter 6, can be expanded into the problem of designing systems that assist and allow users to navigate through without costing them excessive cognitive demands. Tools and methods to quantify the effects of cognitive biases, which we discussed in Chapter 4, can be repurposed to measure the system-imposed cognitive demands. The concept of cognitive bias susceptibility, which we explored in Chapter 5, can be taken into account as a control variable when researching system-imposed cognitive demands because different individuals employ different cognitive strategies to address cognitive demands they experience. Consequently, we envision that the work of this thesis can be generalised into the broader investigation of computing systems imposing cognitive demands on end-users.

This page is intentionally left blank.

Bibliography

[1] Z. Aghajari, E. P. S. Baumer, and D. DiFranzo. Reviewing interventions to address misinformation: The need to expand our vision beyond an individualistic focus. *Proc. ACM Hum.-Comput. Interact.*, 7 (CSCW1), apr 2023. doi: 10.1145/3579520. URL https://doi.org/10.1145/3579520.

[2] F. Alatawi, L. Cheng, A. Tahir, M. Karami, B. Jiang, T. Black, and H. Liu. A survey on echo chambers on social media: Description, detection and mitigation. *arXiv preprint arXiv:2112.05084*, 2021. URL https://arxiv.org/abs/2112.05084.

[3] M. Allais. Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'ecole americaine. *Econometrica*, 21(4):503–546, 1953. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1907921.

[4] G. G. and. How to make cognitive illusions disappear: Beyond "heuristics and biases". *European Review of Social Psychology*, 2(1):83–115, 1991. doi: 10.1080/14792779143000033. URL https://doi.org/10.1080/14792779143000033.

[5] K. Arceneaux. Cognitive biases and the strength of political arguments. *American Journal of Political Science*, 56(2):271–285, 2012. doi: https://doi.org/10.1111/j.1540-5907.2011.00573.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-5907.2011.00573.x.

[6] D. Arnott. A taxonomy of decision biases. *Monash University, School of Information Management and Systems, Caulfield*, 1998.

[7] D. Arnott. Cognitive biases and decision support systems development: a design science approach. *Information Systems Journal*, 16(1):55–78, 2006. doi: https://doi.org/10.1111/j.1365-2575.2006.00208.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2575.2006.00208.x.

[8] L. Azzopardi. Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval. *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pages 27–37, 2021. doi: 10.1145/3406522.3446023. Publisher: Association for Computing Machinery.

[9] E. Babaei, B. Tag, T. Dingler, and E. Velloso. A critique of electrodermal activity practices at chi. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445370. URL https://doi.org/10.1145/3411764.3445370.

[10] A. Baigelenov, P. Shukla, Z. Zhang, and P. Parsons. Are cognitive biases as important as they seem for data visualization? In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713958. doi: 10.1145/3706599.3719926. URL https://doi.org/10.1145/3706599.3719926.

[11] B. Benson. Cognitive bias cheat sheet. https://betterhumans.pub/cognitive-bias-cheat-sheet-55a472476b18, 2016. Accessed: 2024-07-15.

[12] H. Berghel. Malice domestic: The cambridge analytica dystopia. *Computer*, 51(05):84–89, may 2018. ISSN 1558-0814. doi: 10.1109/MC.2018.2381135.

[13] K. Bergram, M. Djokovic, V. Bezençon, and A. Holzer. The digital landscape of nudging: A systematic literature review of empirical research on digital nudges. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517638. URL https://doi.org/10.1145/3491102.3517638.

[14] A. Bertrand, R. Belloum, J. R. Eagan, and W. Maxwell. How cognitive biases affect xai-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 78–91, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392471. doi: 10.1145/3514094.3534164. URL https://doi.org/10.1145/3514094.3534164.

[15] S. Bhadani. Biases in recommendation system. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, page 855–859, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384582. doi: 10.1145/3460231.3473897. URL https://doi.org/10.1145/3460231.3473897.

[16] N. Boonprakong, G. He, U. Gadiraju, N. Van Berkel, D. Wang, S. Chen, J. Liu, B. Tag, J. Goncalves, and T. Dingler. Workshop on understanding and mitigating cognitive biases in human-ai collaboration. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '23 Companion, page 512–517, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701290. doi: 10.1145/3584931.3611284. URL https://doi.org/10.1145/3584931.3611284.

[17] L. Bovens. *The Ethics of Nudge*, pages 207–219. Springer Netherlands, Dordrecht, 2009. ISBN 978-90-481-2593-7. doi: 10.1007/978-90-481-2593-7_10. URL https://doi.org/10.1007/978-90-481-2593-7_10.

[18] M. M. Boyer. Aroused argumentation: How the news exacerbates motivated reasoning. *The International Journal of Press/Politics*, page 19401612211010577, 2021. doi: 10.1177/19401612211010577. URL https://doi.org/10.1177/19401612211010577.

[19] P. Burda, L. Allodi, and N. Zannone. Cognition in Social Engineering Empirical Research: A Systematic Literature Review. *ACM Trans. Comput.-Hum. Interact.*, 31(2), 2024. doi: 10.1145/3635149.

[20] J. T. Cacioppo and R. E. Petty. The need for cognition. *Journal of personality and social psychology*, 42 (1):116, 1982.

[21] S. Cao, A. Liu, and C.-M. Huang. Designing for Appropriate Reliance: The Roles of AI Uncertainty Presentation, Initial User Decision, and User Demographics in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1), 2024. doi: 10.1145/3637318.

[22] A. Caraban, E. Karapanos, D. Gonçalves, and P. Campos. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–15, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300733. URL https://doi.org/10.1145/3290605.3300733.

[23] A. Caraban, E. Karapanos, D. Gonçalves, and P. Campos. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2019. doi: 10.1145/3290605.3300733. Publisher: Association for Computing Machinery.

[24] E. Caragay, K. Xiong, J. Zong, and D. Jackson. Beyond dark patterns: A concept-based framework for ethical software design. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642781. URL https://doi.org/10.1145/3613904.3642781.

[25] S. K. Card, A. Newell, and T. P. Moran. *The Psychology of Human-Computer Interaction*. L. Erlbaum Associates Inc., USA, 1983. ISBN 0898592437.

[26] C. R. Carter, L. Kaufmann, and A. Michel. Behavioral supply management: a taxonomy of judgment and decision-making biases. *International Journal of Physical Distribution & Logistics Management*, 37 (8):631–669, Jan. 2007. ISSN 0960-0035. doi: 10.1108/09600030710825694. URL https://doi.org/10.1108/09600030710825694. Publisher: Emerald Group Publishing Limited.

[27] F. Chiossi, L. Haliburton, C. Ou, A. M. Butz, and A. Schmidt. Short-form videos degrade our capacity to retain intentions: Effect of context switching on prospective memory. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3580778. URL https://doi.org/10.1145/3544548.3580778.

[28] Y.-L. Chou, Y.-H. Lin, T.-Y. Lin, H. Y. You, and Y.-J. Chang. Why did you/i read but not reply? im users' unresponded-to read-receipt practices and explanations of them. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517496. URL https://doi.org/10.1145/3491102.3517496.

[29] A. Cockburn, B. Lewis, P. Quinn, and C. Gutwin. Framing effects influence interface feature decisions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–11, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376496. URL https://doi.org/10.1145/3313831.3376496.

[30] J. Collins. Optimally irrational: The good reasons we behave the way we do, by lionel page (cambridge university press, cambridge, 2022), pp. 322. *Economic Record*, 100(329):271–274, 2024. doi: https://doi.org/10.1111/1475-4932.12798. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-4932.12798.

[31] L. Cosmides. The logic of social exchange: Has natural selection shaped how humans reason? studies with the wason selection task. *Cognition*, 31(3):187–276, 1989. ISSN 0010-0277. doi: https://doi.org/10.1016/0010-0277(89)90023-1. URL https://www.sciencedirect.com/science/article/pii/0010027789900231.

[32] K. A. Costabile and S. B. Klein. Finishing Strong: Recency Effects in Juror Judgments. *Basic and Applied Social Psychology*, 27(1):47–58, 2005. ISSN 1532-4834(Electronic),0197-3533(Print). doi: 10.1207/s15324834basp2701_5. Place: US Publisher: Lawrence Erlbaum.

[33] V. Danry, P. Pataranutaporn, Y. Mao, and P. Maes. Wearable reasoner: Towards enhanced human rationality through a wearable device with an explainable ai assistant. In *Proceedings of the Augmented Humans International Conference*, AHs '20, New York, NY, USA, 2020. Association for Computing Machin-

ery. ISBN 9781450376037. doi: 10.1145/3384657.3384799. URL https://doi.org/10.1145/3384657.3384799.

[34] D. de Ridder, F. Kroese, and L. van Gestel. Nudgeability: Mapping conditions of susceptibility to nudge influence. *Perspectives on Psychological Science*, 17(2):346–359, 2022. doi: 10.1177/1745691621995183. URL https://doi.org/10.1177/1745691621995183. PMID: 34424801.

[35] T. Dingler, B. Tag, E. Karapanos, K. Kise, and A. Dengel. Workshop on detection and design for cognitive biases in people and computing systems. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–6, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368193. doi: 10.1145/3334480.3375159. URL https://doi.org/10.1145/3334480.3375159.

[36] T. Dingler, B. Tag, P. Lorenz-Spreen, A. W. Vargo, S. Knight, and S. Lewandowsky. Workshop on Technologies to Support Critical Thinking in an Age of Misinformation. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–5, New York, NY, USA, may 2021. ACM. ISBN 9781450380959. doi: 10.1145/3411763.3441350. URL https://dl.acm.org/doi/10.1145/3411763.3441350.

[37] T. Dingler, B. Tag, and A. Vargo. Technologies to Support Critical Thinking in an Age of Misinformation (Dagstuhl Seminar 22172). *Dagstuhl Reports*, 12(4):72–95, 2022. ISSN 2192-5283. doi: 10.4230/DagRep.12.4.72. URL https://drops.dagstuhl.de/opus/volltexte/2022/17281.

[38] S. Doroudi and S. A. Rastegar. The bias–variance tradeoff in cognitive science. *Cognitive Science*, 47(1):e13241, 2023. doi: https://doi.org/10.1111/cogs.13241. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.13241.

[39] A. Downey. *Think Bayes.* " O'Reilly Media, Inc.", 2021.

[40] K. Edwards and E. E. Smith. A disconfirmation bias in the evaluation of arguments. *Journal of Personality and Social Psychology*, 71(1):5–24, 1996. ISSN 1939-1315(Electronic),0022-3514(Print). doi: 10.1037/0022-3514.71.1.5. Place: US Publisher: American Psychological Association.

[41] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, and J. D. Weisz. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445188. URL https://doi.org/10.1145/3411764.3445188.

[42] M. I. Eronen and L. F. Bringmann. The Theory Crisis in Psychology: How to Move Forward. *Perspectives on psychological science : a journal of the Association for Psychological Science*, 16(4):779–788, July 2021. ISSN 1745-6924 1745-6916. doi: 10.1177/1745691620970586. Place: United States.

[43] J. S. B. Evans. *Bias in human reasoning: Causes and consequences.* Lawrence Erlbaum Associates, Inc, 1989.

[44] J. S. B. T. Evans. Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 59(1):255–278, 2008. doi: 10.1146/annurev.psych.59.103006.093629. URL https://doi.org/10.1146/annurev.psych.59.103006.093629. PMID: 18154502.

[45] J. S. B. T. Evans. Reasoning, biases and dual processes: The lasting impact of wason (1960). *Quarterly Journal of Experimental Psychology*, 69(10):2076–2092, 2016. doi: 10.1080/17470218.2014.914547. URL https://doi.org/10.1080/17470218.2014.914547. PMID: 25158629.

[46] J. S. B. T. Evans and K. E. Stanovich. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3):223–241, 2013. doi: 10.1177/1745691612460685. URL https://doi.org/10.1177/1745691612460685. PMID: 26172965.

[47] A. T. Fiore and J. S. Donath. Homophily in online dating: when do you like someone like yourself? In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '05, page 1371–1374, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595930027. doi: 10.1145/1056808.1056919. URL https://doi.org/10.1145/1056808.1056919.

[48] C. B. Foundation. Bias cheat sheet. http://bias.transhumanity.net/bias-cheat-sheet/, 2024. Accessed: 2024-09-13.

[49] S. Frederick. Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4):25–42, December 2005. doi: 10.1257/089533005775196732. URL https://www.aeaweb.org/articles?id=10.1257/089533005775196732.

[50] J. J. Gibson. *The ecological approach to visual perception*. Houghton Mifflin Harcourt, Boston, 1979. ISBN 0205093103. doi: 10.5962/pb.gibson.1979.

[51] G. Gigerenzer. Ecological intelligence: An adaptation for frequencies. In *The evolution of mind.*, pages 9–29. Oxford University Press, New York, NY, US, 1998. ISBN 0-19-511053-6 (Hardcover).

[52] G. Gigerenzer. *Fast and Frugal Heuristics: The Tools of Bounded Rationality*, chapter 4, pages 62–88. John Wiley & Sons, Ltd, 2004. ISBN 9780470752937. doi: https://doi.org/10.1002/9780470752937.ch4. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470752937.ch4.

[53] G. Gigerenzer. The rationality wars: A personal reflection. *Behavioural Public Policy*, page 1–21, 2024. doi: 10.1017/bpp.2024.51.

[54] G. Gigerenzer and H. Brighton. Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, 1(1):107–143, 2009. doi: https://doi.org/10.1111/j.1756-8765.2008.01006.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1756-8765.2008.01006.x.

[55] G. Gigerenzer and D. G. Goldstein. Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, 103(4):650–669, Oct. 1996. ISSN 0033-295X. doi: 10.1037/0033-295x.103.4.650. Place: United States.

[56] G. Gigerenzer and U. Hoffrage. How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102(4):684–704, 1995. ISSN 1939-1471(Electronic),0033-295X(Print). doi: 10.1037/0033-295X.102.4.684. Place: US Publisher: American Psychological Association.

[57] G. Gigerenzer and P. M. Todd. *Simple heuristics that make us smart.* Simple heuristics that make us smart. Oxford University Press, New York, NY, US, 1999. ISBN 0-19-512156-2 (Hardcover). Pages: xv, 416.

[58] G. Gigerenzer, P. M. Todd, and A. B. C. R. Group. *Simple Heuristics That Make Us Smart.* Oxford University Press USA, New York, NY, USA, 1999.

[59] S. J. J. Gould, L. L. Chuang, I. Iacovides, D. Garaialde, M. E. Cecchinato, B. R. Cowan, and A. L. Cox. A special interest group on designed and engineered friction in interaction. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380959. doi: 10.1145/3411763.3450404. URL https://doi.org/10.1145/3411763.3450404.

[60] N. Goyal, C. Baumler, T. Nguyen, and H. Daumé III. The impact of explanations on fairness in human-ai decision-making: Protected vs proxy features. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, page 155–180, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705083. doi: 10.1145/3640543.3645210. URL https://doi.org/10.1145/3640543.3645210.

[61] E. Graells-Garrido, M. Lalmas, and R. Baeza-Yates. Data Portraits and Intermediary Topics: Encouraging Exploration of Politically Diverse Profiles. *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 228–240, 2016. doi: 10.1145/2856767.2856776. Publisher: Association for Computing Machinery.

[62] C. M. Gray, Y. Kou, B. Battles, J. Hoggatt, and A. L. Toombs. The dark (patterns) side of ux design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–14, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3174108. URL https://doi.org/10.1145/3173574.3174108.

[63] A. G. Greenwald, D. E. McGhee, and J. L. K. Schwartz. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74 6:1464–80, 1998. URL https://api.semanticscholar.org/CorpusID:7840819.

[64] X. Guo, L. Wang, M. Zhang, and G. Chen. First Things First? Order Effects in Online Product Recommender Systems. *ACM Trans. Comput.-Hum. Interact.*, 30(1), 2023. doi: 10.1145/3557886.

[65] M. S. Hagger. Avoiding the "déjà-variable" phenomenon: social psychology needs more guides to constructs. *Frontiers in psychology*, 5:52, 2014. ISSN 1664-1078. doi: 10.3389/fpsyg.2014.00052. Place: Switzerland.

[66] P. G. Hansen and A. M. Jespersen. Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy. *European Journal of Risk Regulation*, 4(1):3–28, 2013. doi: 10.1017/S1867299X00002762.

[67] J. Hartmann, A. De Angeli, and A. Sutcliffe. Framing the user experience: information biases on website quality judgement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, page 855–864, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605580111. doi: 10.1145/1357054.1357190. URL https://doi.org/10.1145/1357054.1357190.

[68] M. G. Haselton and D. M. Buss. Error management theory and the evolution of misbeliefs. *Behavioral and Brain Sciences*, 32(6):522–523, 2009. doi: 10.1017/S0140525X09991440.

[69] M. G. Haselton, G. A. Bryant, A. Wilke, D. A. Frederick, A. Galperin, W. E. Frankenhuis, and T. Moore. Adaptive rationality: An evolutionary perspective on cognitive bias. *Social Cognition*, 27(5):733–763, 2009. doi: 10.1521/soco.2009.27.5.733. URL https://doi.org/10.1521/soco.2009.27.5.733.

[70] M. G. Haselton, D. Nettle, and D. R. Murray. The evolution of cognitive bias. *The handbook of evolutionary psychology*, pages 1–20, 2015.

[71] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

[72] E. B. Hekler, P. Klasnja, J. E. Froehlich, and M. P. Buman. Mind the Theoretical Gap: Interpreting, Using, and Developing Behavioral Theory in HCI Research. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3307–3316, 2013. doi: 10.1145/2470654.2466452. Publisher: Association for Computing Machinery.

[73] J. Henrich and F. J. Gil-White. The evolution of prestige: freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior*, 22(3): 165–196, 2001. ISSN 1090-5138. doi: https://doi.org/10.1016/S1090-5138(00)00071-4. URL https://www.sciencedirect.com/science/article/pii/S1090513800000714.

[74] R. Hertwig, J. Nerissa Davis, and F. J. Sulloway. Parental investment: How an equity motive can produce inequality. In *Heuristics: The Foundations of Adaptive Behavior*. Oxford University Press, 04 2011. ISBN 9780199744282. doi: 10.1093/acprof:oso/9780199744282.003.0035. URL https://doi.org/10.1093/acprof:oso/9780199744282.003.0035.

[75] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk. Session-based recommendations with recurrent neural networks, 2016. URL https://arxiv.org/abs/1511.06939.

[76] U. Hoffrage, S. Lindsey, R. Hertwig, and G. Gigerenzer. Communicating statistical information. *Science*, 290(5500):2261–2262, 2000. doi: 10.1126/science.290.5500.2261. URL https://www.science.org/doi/abs/10.1126/science.290.5500.2261.

[77] K. Hornbæk, A. Mottelson, J. Knibbe, and D. Vogel. What do we mean by "interaction"? an analysis of 35 years of chi. *ACM Trans. Comput.-Hum. Interact.*, 26(4), jul 2019. ISSN 1073-0516. doi: 10.1145/3325285. URL https://doi.org/10.1145/3325285.

[78] C. K. Hsee. The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes*, 67(3): 247–257, 1996. ISSN 0749-5978. doi: https://doi.org/10.1006/obhd.1996.0077. URL https://www.sciencedirect.com/science/article/pii/S0749597896900771.

[79] A. Jain, A. H. Horowitz, F. Schoeller, S.-w. Leigh, P. Maes, and M. Sra. Designing interactions beyond conscious control: A new model for wearable interfaces. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(3), sep 2020. doi: 10.1145/3411829. URL https://doi.org/10.1145/3411829.

[80] W. James. *The Principles of Psychology*, volume 1. Macmillan London, 1890.

[81] K. Ji, D. Spina, D. Hettiachchi, F. D. Salim, and F. Scholer. Examining the impact of uncontrolled variables on physiological signals in user studies for information processing activities. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 1971–1975, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394086. doi: 10.1145/3539618.3591981. URL https://doi.org/10.1145/3539618.3591981.

[82] K. Ji, S. Pathiyan Cherumanal, J. R. Trippas, D. Hettiachchi, F. D. Salim, F. Scholer, and D. Spina. Towards detecting and mitigating cognitive bias in spoken conversational search. In *Adjunct Proceedings of the 26th International Conference on Mobile Human-Computer Interaction*, MobileHCI '24 Adjunct, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705069. doi: 10.1145/3640471.3680245. URL https://doi.org/10.1145/3640471.3680245.

[83] P. N. Johnson-Laird. *Mental models: towards a cognitive science of language, inference, and consciousness*. Harvard University Press, USA, 1986. ISBN 0674568826.

[84] C. Jolls and C. R. Sunstein. Debiasing through law. *The Journal of Legal Studies*, 35(1):199–242, 2006. doi: 10.1086/500096. URL https://doi.org/10.1086/500096.

[85] A. R. Jr and K. Broneck. 'im me': Instant messaging as relational maintenance and everyday communication. *Journal of Social and Personal Relationships*, 26(2-3):291–314, 2009. doi: 10.1177/0265407509106719. URL https://doi.org/10.1177/0265407509106719.

[86] D. Kahneman. Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review*, 93(5):1449–1475, 2003. ISSN 00028282. URL http://www.jstor.org/stable/3132137.

[87] D. Kahneman. *Thinking, Fast and Slow*. Macmillan, 2011.

[88] D. Kahneman. A new etiquette for replication. *Social Psychology*, 45(4):310–311, 2014. ISSN 2151-2590(Electronic),1864-9335(Print). Place: Germany Publisher: Hogrefe Publishing.

[89] D. Kahneman and A. Tversky. Subjective probability: A judgment of representativeness. *Cognitive psychology*, 3(3):430–454, 1972.

[90] D. Kahneman and A. Tversky. On the psychology of prediction. *Psychological Review*, 80 (4):237 – 251, 1973. doi: 10.1037/h0034747. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-58149417364&doi=10.1037%2fh0034747&partnerID=40&md5=b44cd7a883920196cde40ff1e1eb5445. Cited by: 3720.

[91] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47 (2):263–291, 1979. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1914185.

[92] D. Kahneman, P. Slovic, and A. Tversky. *Judgment Under Uncertainty: Heuristics and biases*. Cambridge university press, 1982.

[93] D. Kahneman, B. L. Fredrickson, C. A. Schreiber, and D. A. Redelmeier. When more pain is preferred to less: Adding a better end. *Psychological Science*, 4(6):401–405, 1993. ISSN 09567976, 14679280. URL http://www.jstor.org/stable/40062570.

[94] D. Kahneman, O. Sibony, and C. R. Sunstein. *Noise: A flaw in human judgment*. Hachette UK, 2021.

[95] H. Kaplan. A theory of fertility and parental investment in traditional and modern human societies. *American Journal of Physical Anthropology*, 101(S23):91–135, 1996. doi: https://doi.org/10.1002/(SICI)1096-8644(1996)23+⟨91::AID-AJPA4⟩3.0.CO;2-C. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291096-8644%281996%2923%2B%3C91%3A%3AAID-AJPA4%3E3.0.CO%3B2-C.

[96] J. T. Kaplan, S. I. Gimbel, and S. Harris. Neural correlates of maintaining one's political beliefs in the face of counterevidence. *Scientific reports*, 6:39589, December 2016. ISSN 2045-2322. doi: 10.1038/srep39589. URL https://europepmc.org/articles/PMC5180221.

[97] H. Kaur, M. R. Conrad, D. Rule, C. Lampe, and E. Gilbert. Interpretability gone bad: The role of bounded rationality in how practitioners understand machine learning. *Proc. ACM Hum.-Comput. Interact.*, 8 (CSCW1), apr 2024. doi: 10.1145/3637354. URL https://doi.org/10.1145/3637354.

[98] G. Keil and N. Kreft. *Human Beings as Rational Animals*, page 23–96. Cambridge University Press, 2019.

[99] S. S. Y. Kim, J. W. Vaughan, Q. V. Liao, T. Lombrozo, and O. Russakovsky. Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3714020. URL https://doi.org/10.1145/3706598.3714020.

[100] P. A. Klaczynski. Motivated scientific reasoning biases, epistemological beliefs, and theory polarization: A two-process approach to adolescent cognition. *Child Development*, 71(5):1347–1366, 2000. doi: https://doi.org/10.1111/1467-8624.00232.

[101] T. Kliegr, Štěpán Bahník, and J. Fürnkranz. A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence*, 295:103458, 2021. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2021.103458. URL https://www.sciencedirect.com/science/article/pii/S0004370221000096.

[102] H.-K. Kong, Z. Liu, and K. Karahalios. Trust and Recall of Information across Varying Degrees of Title-Visualization Misalignment. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019. doi: 10.1145/3290605.3300576. Publisher: Association for Computing Machinery.

[103] J. A. Krosnick, C. M. Judd, and B. Wittenbrink. The measurement of attitudes. *The handbook of attitudes*, 21:76, 2005.

[104] J. Kruger. Lake wobegon be gone! the" below-average effect" and the egocentric nature of comparative ability judgments. *Journal of personality and social psychology*, 77(2):221, 1999.

[105] P. Kuyer and B. Gordijn. Nudge in perspective: A systematic literature review on the ethical issues with nudging. *Rationality and Society*, 35(2):191–230, 2023. doi: 10.1177/10434631231155005. URL https://doi.org/10.1177/10434631231155005.

[106] N.-T. Le and L. Wartschinski. A Cognitive Assistant for improving human reasoning skills. *International Journal of Human-Computer Studies*, 117:45–54, 2018. doi: 10.1016/j.ijhcs.2018.02.005.

[107] A. R. Lee, S.-M. Son, and K. K. Kim. Information and communication technology overload and social networking service fatigue: A stress perspective. *Computers in Human Behavior*, 55:51–61, 2016. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2015.08.011. URL https://www.sciencedirect.com/science/article/pii/S0747563215300893.

[108] S. Lewis, M. Dontcheva, and E. Gerber. Affective computational priming and creativity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 735–744, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450302289. doi: 10.1145/1978942.1979048. URL https://doi.org/10.1145/1978942.1979048.

[109] J. Li, S. Zhang, and W. Ao. Why is instant messaging not instant? understanding users' negative use behavior of instant messaging software. *Computers in Human Behavior*, 142:107655, 2023. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2023.107655. URL https://www.sciencedirect.com/science/article/pii/S0747563223000067.

[110] Q. V. Liao and W.-T. Fu. Beyond the filter bubble: Interactive effects of perceived threat and topic involvement on selective exposure to information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 2359–2368, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450318990. doi: 10.1145/2470654.2481326. URL https://doi.org/10.1145/2470654.2481326.

[111] S. O. Lilienfeld, R. Ammirati, and K. Landfield. Giving debiasing away: Can psychological research on correcting cognitive errors promote human welfare? *Perspectives on Psychological Science*, 4(4):390–398, 2009. doi: 10.1111/j.1745-6924.2009.01144.x. URL https://doi.org/10.1111/j.1745-6924.2009.01144.x. PMID: 26158987.

[112] J. Liu. Toward A Two-Sided Fairness Framework in Search and Recommendation. *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, pages 236–246, 2023. doi: 10.1145/3576840.3578332. Publisher: Association for Computing Machinery.

[113] Q. Liu, H. Jiang, Z. Pan, Q. Han, Z. Peng, and Q. Li. BiasEye: A Bias-Aware Real-time Interactive Material Screening System for Impartial Candidate Assessment. *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 325–343, 2024. doi: 10.1145/3640543.3645166. Publisher: Association for Computing Machinery.

[114] Z. Ma and K. Z. Gajos. Not just a preference: Reducing biased decision-making on dating websites. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517587. URL https://doi.org/10.1145/3491102.3517587.

[115] P. G. Mahon and R. L. Canosa. Prisoners and chickens: gaze locations indicate bounded rationality. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '12, page 401–404, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450312219. doi: 10.1145/2168556.2168647. URL https://doi.org/10.1145/2168556.2168647.

[116] A. Mathur, G. Acar, M. J. Friedman, E. Lucherini, J. Mayer, M. Chetty, and A. Narayanan. Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), 2019. doi: 10.1145/3359183.

[117] A. Mathur, M. Kshirsagar, and J. Mayer. What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021. doi: 10.1145/3411764.3445610. Publisher: Association for Computing Machinery.

[118] S. E. Maxwell, M. Y. Lau, and G. S. Howard. Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, 70(6):487–498, 2015. ISSN 1935-990X(Electronic),0003-066X(Print). doi: 10.1037/a0039400. Place: US Publisher: American Psychological Association.

[119] S. Mertens, M. Herberz, U. J. J. Hahnel, and T. Brosch. The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proceedings of the National Academy of Sciences*, 119(1):e2107346118, 2022. doi: 10.1073/pnas.2107346118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2107346118.

[120] W. S. Messer and R. A. Griggs. Another look at Linda. *Bulletin of the Psychonomic Society*, 31(3):193–196, Mar. 1993. ISSN 0090-5054. doi: 10.3758/BF03337322. URL https://doi.org/10.3758/BF03337322.

[121] J. S. Mill. On the definition and method of political economy. *The philosophy of economics*, pages 41–58, 1836.

[122] D. Moher, A. Liberati, J. Tetzlaff, and D. Altman. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Br Med J*, 8:336–341, 07 2009. doi: 10.1371/joumal.pmedl000097.

[123] Z. Munn, M. D. J. Peters, C. Stern, C. Tufanaru, A. McArthur, and E. Aromataris. Systematic review or scoping review? guidance for authors when choosing between a systematic or scoping review approach. *BMC Med. Res. Methodol.*, 18(1):143, Nov. 2018.

[124] R. S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

[125] D. A. Norman. *The Design of Everyday Things*. Basic Books, Inc., USA, 2002. ISBN 9780465067107.

[126] M. I. Norton, D. Mochon, and D. Ariely. The ikea effect: When labor leads to love. *Journal of Consumer Psychology*, 22(3):453–460, 2012. ISSN 1057-7408. doi: https://doi.org/10.1016/j.jcps.2011.08.002. URL https://www.sciencedirect.com/science/article/pii/S1057740811000829.

[127] M. Nourani, C. Roy, J. E. Block, D. R. Honeycutt, T. Rahman, E. Ragan, and V. Gogate. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. *26th International Conference on Intelligent User Interfaces*, pages 340–350, 2021. doi: 10.1145/3397481.3450639. Publisher: Association for Computing Machinery.

[128] A. Oeberst and R. Imhoff. Toward parsimony in bias research: A proposed common framework of belief-consistent information processing for a set of biases. *Perspectives on Psychological Science*, 18(6):1464–1487, 2023. doi: 10.1177/17456916221148147. URL https://doi.org/10.1177/17456916221148147. PMID: 36930530.

[129] T. Okoshi, W. Sasaki, and J. Nakazawa. Behavification: bypassing human's attentional and cognitive systems for automated behavior change. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, UbiComp/ISWC '20 Adjunct, page 692–695, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380768. doi: 10.1145/3410530.3414439. URL https://doi.org/10.1145/3410530.3414439.

[130] S. Oreg and M. Bayazit. Prone to bias: Development of a bias taxonomy from an individual differences perspective. *Review of General Psychology*, 13(3):175–193, 2009. doi: 10.1037/a0015656. URL https://doi.org/10.1037/a0015656.

[131] L. Page. *Optimally Irrational: The Good Reasons We Behave the Way We Do*. Cambridge University Press, 2022. ISBN 9781009209205.

[132] L. Page. Adaptive explanations to behavioural findings. In *Elgar Encyclopedia of Behavioural and Experimental Economics*, pages 1–3. Edward Elgar Publishing, Feb. 2025. doi: https://doi.org/10.4337/9781802207736.00007.

[133] G. Payne and J. Payne. *Key Concepts in Social Research*. SAGE Publications, Ltd, London, 2004. doi: 10.4135/9781849209397. URL https://methods.sagepub.com/dict/mono/key-concepts-in-social-research/toc.

[134] G. Phillips-Wren and M. Adya. Decision making under stress: the role of information overload, time pressure, complexity, and uncertainty. *Journal of Decision Systems*, 29(sup1):213–225, 2020. doi: 10.1080/12460125.2020.1768680. URL https://doi.org/10.1080/12460125.2020.1768680.

[135] G. W. Ploger, J. Dunaway, P. Fournier, and S. Soroka. The psychophysiological correlates of cognitive dissonance. *Politics and the Life Sciences*, 40(2):202–212, 2021. doi: 10.1017/pls.2021.15.

[136] R. Pohl. *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*. Psychology Press, 2004.

[137] S. Pothirattanachaikul, T. Yamamoto, Y. Yamamoto, and M. Yoshikawa. Analyzing the Effects of "People Also Ask" on Search Behaviors and Beliefs. *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, pages 101–110, 2020. doi: 10.1145/3372923.3404786. Publisher: Association for Computing Machinery.

[138] T. Richter. Validation and comprehension of text information: Two sides of the same coin. *Discourse Processes*, 52(5-6):337–355, 2015. doi: 10.1080/0163853X.2015.1025665.

[139] T. Richter and J. Maier. Comprehension of multiple documents with conflicting information: A two-step model of validation. *Educational Psychologist*, 52(3):148–166, 2017. doi: 10.1080/00461520.2017.1322968.

[140] A. Rieger, T. Draws, M. Theune, and N. Tintarev. This item might reinforce your opinion: Obfuscation and labeling of search results to mitigate confirmation bias. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, HT '21, page 189–199, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385510. doi: 10.1145/3465336.3475101. URL https://doi.org/10.1145/3465336.3475101.

[141] A. Rieger, T. Draws, M. Theune, and N. Tintarev. This Item Might Reinforce Your Opinion: Obfuscation and Labeling of Search Results to Mitigate Confirmation Bias. *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pages 189–199, 2021. doi: 10.1145/3465336.3475101. Publisher: Association for Computing Machinery.

[142] A. Rieger, Q.-U.-A. Shaheen, C. Sierra, M. Theune, and N. Tintarev. Towards Healthy Engagement with Online Debates: An Investigation of Debate Summaries and Personalized Persuasive Suggestions. *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, pages 192–199, 2022. doi: 10.1145/3511047.3537692. Publisher: Association for Computing Machinery.

[143] A. Rieger, F. Bredius, N. Tintarev, and M. S. Pera. 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*, CHI EA '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394222. doi: 10.1145/3544549.3585693. URL https://doi.org/10.1145/3544549.3585693.

[144] J. Rieskamp and A. Dieckmann. Redundancy: Environment structure that simple heuristics can exploit. In *Ecological Rationality: Intelligence in the World*. Oxford University Press, 03 2012. ISBN 9780195315448. doi: 10.1093/acprof:oso/9780195315448.003.0056. URL https://doi.org/10.1093/acprof:oso/9780195315448.003.0056.

[145] Y. Rogers. *HCI Theory: Classical, Modern, and Contemporary*. Morgan & Claypool Publishers, 1st edition, 2012. ISBN 1608459004.

[146] L. Ross. The intuitive psychologist and his shortcomings: Distortions in the attribution process. volume 10 of *Advances in Experimental Social Psychology*, pages 173–220. Academic Press, 1977. doi: https://doi.org/10.1016/S0065-2601(08)60357-3. URL https://www.sciencedirect.com/science/article/pii/S0065260108603573.

[147] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x. URL https://doi.org/10.1038/s42256-019-0048-x.

[148] J. Sabini, M. Siepmann, and J. Stein. The really fundamental attribution error in social psychological research. *Psychological Inquiry*, 12(1):1–15, 2001. ISSN 1047840X, 15327965. URL http://www.jstor.org/stable/1449294.

[149] R. Samuels, S. Stich, and M. Bishop. Ending the rationality wars how to make disputes about human rationality disappear. In *Common Sense, Reasoning, and Rationality*. Oxford University Press, 02 2002. ISBN 9780195147667. doi: 10.1093/0195147669.003.0011. URL https://doi.org/10.1093/0195147669.003.0011.

[150] S. Santhanam, A. Karduni, and S. Shaikh. Studying the effects of cognitive biases in evaluation of conversational agents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376318. URL https://doi.org/10.1145/3313831.3376318.

[151] G. Saposnik, D. Redelmeier, C. C. Ruff, and P. N. Tobler. Cognitive biases associated with medical decisions: a systematic review. *BMC Medical Informatics and Decision Making*, 16(1):138, Nov. 2016. ISSN 1472-6947. doi: 10.1186/s12911-016-0377-1. URL https://doi.org/10.1186/s12911-016-0377-1.

[152] C. Schwind, J. Buder, and F. W. Hesse. I Will Do It, but i Don't like It: User Reactions to Preference-Inconsistent Recommendations. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 349–352, 2011. doi: 10.1145/1978942.1978992. Publisher: Association for Computing Machinery.

[153] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 59–68, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287598. URL https://doi.org/10.1145/3287560.3287598.

[154] N. Sharma, Q. V. Liao, and Z. Xiao. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024. doi: 10.1145/3613904.3642459. Publisher: Association for Computing Machinery.

[155] R. J. Shiller. From efficient markets theory to behavioral finance. *Journal of Economic Perspectives*, 17(1): 83–104, March 2003. doi: 10.1257/089533003321164967. URL https://www.aeaweb.org/articles?id=10.1257/089533003321164967.

[156] H. A. Simon. *Administrative behavior; a study of decision-making processes in administrative organization.* Administrative behavior; a study of decision-making processes in administrative organization. Macmillan, Oxford, England, 1947. Pages: xvi, 259.

[157] H. A. Simon. A behavioral model of rational choice. *Models of man, social and rational: Mathematical essays on rational human behavior in a social setting*, pages 241–260, 1957.

[158] H. A. Simon. *From substantive to procedural rationality*, pages 65–86. Springer US, Boston, MA, 1976. ISBN 978-1-4613-4367-7. doi: 10.1007/978-1-4613-4367-7_6. URL https://doi.org/10.1007/978-1-4613-4367-7_6.

[159] H. A. Simon. *Bounded Rationality*, pages 15–18. Palgrave Macmillan UK, London, 1990. ISBN 978-1-349-20568-4. doi: 10.1007/978-1-349-20568-4_5. URL https://doi.org/10.1007/978-1-349-20568-4_5.

[160] H. A. Simon. Invariants of human behavior. *Annual Review of Psychology*, 41(Volume 41, 1990):1–20, 1990. ISSN 1545-2085. doi: https://doi.org/10.1146/annurev.ps.41.020190.000245. URL https://www.annualreviews.org/content/journals/10.1146/annurev.ps.41.020190.000245.

[161] H. A. Simon and A. Newell. Heuristic problem solving: The next advance in operations research. *Operations Research*, 6(1):1–10, 1958. doi: 10.1287/opre.6.1.1. URL https://doi.org/10.1287/opre.6.1.1.

[162] L. J. SKITKA, K. L. MOSIER, and M. BURDICK. Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991–1006, 1999. ISSN 1071-5819. doi: https://doi.org/10.1006/ijhc.1999.0252. URL https://www.sciencedirect.com/science/article/pii/S1071581999902525.

[163] M. Sohn, T. Nam, and W. Lee. Designing with unconscious human behaviors for eco-friendly interaction. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '09, page 2651–2654, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605582474. doi: 10.1145/1520340.1520375. URL https://doi.org/10.1145/1520340.1520375.

[164] P. Srinivasan. "am i overwhelmed with this information?": A cross-platform study on information overload, technostress, well-being, and continued social media usage intentions. In *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '20 Companion, page 165–170, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380591. doi: 10.1145/3406865.3418371. URL https://doi.org/10.1145/3406865.3418371.

[165] N. Srivastava, J. Healey, R. Jain, G. Liu, Y. Ma, B. Llana, D. Gasevic, T. Dingler, and S. Wallace. Priming at scale: An evaluation of using ai to generate primes for mobile readers. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713958. doi: 10.1145/3706599.3720153. URL https://doi.org/10.1145/3706599.3720153.

[166] K. E. Stanovich. *Who is Rational?: Studies of Individual Differences in Reasoning*. Psychology Press, 1999.

[167] K. E. Stanovich and R. F. West. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23(5):645–665, 2000. doi: 10.1017/S0140525X00003435.

[168] E. Stefanidi, M. Bentvelzen, P. W. Woźniak, T. Kosch, M. P. Woźniak, T. Mildner, S. Schneegass, H. Müller, and J. Niess. Literature reviews in hci: A review of reviews. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581332. URL https://doi.org/10.1145/3544548.3581332.

[169] C. R. Sunstein. Nudging and choice architecture: Ethical considerations. *Yale Journal on Regulation, Forthcoming*, 2015.

[170] C. R. Sunstein and R. H. Thaler. Libertarian paternalism is not an oxymoron. *The University of Chicago Law Review*, 70(4):1159–1202, 2003. ISSN 00419494. URL http://www.jstor.org/stable/1600573.

[171] L. Tankelevitch, V. Kewenig, A. Simkute, A. E. Scott, A. Sarkar, A. Sellen, and S. Rintel. The metacognitive demands and opportunities of generative ai. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642902. URL https://doi.org/10.1145/3613904.3642902.

[172] R. Thaler. Toward a positive theory of consumer choice. *Journal of Economic Behavior Organization*, 1(1):39–60, 1980. ISSN 0167-2681. doi: https://doi.org/10.1016/0167-2681(80)90051-7. URL https://www.sciencedirect.com/science/article/pii/0167268180900517.

[173] R. H. Thaler. From homo economicus to homo sapiens. *Journal of Economic Perspectives*, 14(1):133–141, March 2000. doi: 10.1257/jep.14.1.133. URL https://www.aeaweb.org/articles?id=10.1257/jep.14.1.133.

[174] R. H. Thaler. Nudge, not sludge. *Science*, 361(6401):431–431, 2018. doi: 10.1126/science.aau9241. URL https://www.science.org/doi/abs/10.1126/science.aau9241.

[175] R. H. Thaler and C. R. Sunstein. Nudge: Improving decisions about health, wealth, and happiness. *Nudge: Improving decisions about health, wealth, and happiness.*, pages x, 293–x, 293, 2008. ISSN 978-0-300-12223-7 (Hardcover). Place: New Haven, CT, US Publisher: Yale University Press.

[176] G. Theocharous, J. Healey, S. Mahadevan, and M. Saad. Personalizing with Human Cognitive Biases. *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 13–17, 2019. doi: 10.1145/3314183.3323453. Publisher: Association for Computing Machinery.

[177] A. Tversky and D. Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2):207–232, 1973. ISSN 0010-0285. doi: https://doi.org/10.1016/0010-0285(73)90033-9. URL https://www.sciencedirect.com/science/article/pii/0010028573900339.

[178] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157): 1124–1131, 1974.

[179] A. Tversky and D. Kahneman. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4):293–315, 1983. ISSN 1939-1471(Electronic),0033-295X(Print). doi: 10.1037/0033-295X.90.4.293. Place: US Publisher: American Psychological Association.

[180] A. Tversky and D. Kahneman. Rational choice and the framing of decisions. *Decision making: Descriptive, normative, and prescriptive interactions*, pages 167–192, 1988.

[181] A. Tversky and D. Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, Oct. 1992. ISSN 1573-0476. doi: 10.1007/BF00122574. URL https://doi.org/10.1007/BF00122574.

[182] K. Vaccaro, D. Huang, M. Eslami, C. Sandvig, K. Hamilton, and K. Karahalios. The illusion of control: Placebo effects of control settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3173590. URL https://doi.org/10.1145/3173574.3173590.

[183] D. Vedejová and V. Čavojová. Confirmation bias in information search, interpretation, and memory recall: evidence from reasoning about four controversial topics. *Thinking & Reasoning*, 28(1):1–28, 2022. doi: 10.1080/13546783.2021.1891967. URL https://doi.org/10.1080/13546783.2021.1891967.

[184] E. S. Veinott, J. Leonard, E. L. Papautsky, B. Perelman, A. Stankovic, J. Lorince, J. Hotaling, T. Ross, P. Todd, E. Castronova, J. Busemeyer, C. Hale, R. Catrambone, E. Whitaker, O. Fox, J. Flach, and R. R. Hoffman. The effect of camera perspective and session duration on training decision making in a serious video game. *2013 IEEE International Games Innovation Conference (IGIC)*, pages 256–262, 2013. doi: 10.1109/IGIC.2013.6659170.

[185] N. Voigtländer and H.-J. Voth. Nazi indoctrination and anti-semitic beliefs in germany. *Proceedings of the National Academy of Sciences*, 112(26):7931–7936, 2015. doi: 10.1073/pnas.1414822112. URL https://www.pnas.org/doi/abs/10.1073/pnas.1414822112.

[186] J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Theory of games and economic behavior. Princeton University Press, Princeton, NJ, US, 1944.

[187] P. B. Vranas. Gigerenzer's normative critique of kahneman and tversky. *Cognition*, 76(3):179–193, 2000. ISSN 0010-0277. doi: https://doi.org/10.1016/S0010-0277(99)00084-0. URL https://www.sciencedirect.com/science/article/pii/S0010027799000840.

[188] S. Wadinambiarachchi, R. M. Kelly, S. Pareek, Q. Zhou, and E. Velloso. The effects of generative ai on design fixation and divergent thinking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642919. URL https://doi.org/10.1145/3613904.3642919.

[189] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim. Designing Theory-Driven User-Centric Explainable AI. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2019. doi: 10.1145/3290605.3300831. Publisher: Association for Computing Machinery.

[190] P. Wason and J. Evans. Dual processes in reasoning? *Cognition*, 3(2):141–154, 1974. ISSN 0010-0277. doi: https://doi.org/10.1016/0010-0277(74)90017-1. URL https://www.sciencedirect.com/science/article/pii/0010027774900171.

[191] P. C. Wason. On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12(3):129–140, 1960. doi: 10.1080/17470216008416717. URL https://doi.org/10.1080/17470216008416717.

[192] P. C. Wason. Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20(3):273–281, 1968. doi: 10.1080/14640746808400161. URL https://doi.org/10.1080/14640746808400161.

[193] L. Weidinger, M. Rauh, N. Marchal, A. Manzini, L. A. Hendricks, J. Mateos-Garcia, S. Bergman, J. Kay, C. Griffin, B. Bariach, I. Gabriel, V. Rieser, and W. Isaac. Sociotechnical safety evaluation of generative ai systems, 2023. URL https://arxiv.org/abs/2310.11986.

[194] I. Weinstein. Don't believe everything you think: Cognitive bias in legal decision making. *Clinical Law Review*, 8:783, 2003. URL https://ssrn.com/abstract=2779670. Fordham Law Legal Studies Research Paper No. 2779670, Available at SSRN: https://ssrn.com/abstract=2779670.

[195] D. Westen, P. S. Blagov, K. Harenski, C. Kilts, and S. Hamann. Neural Bases of Motivated Reasoning: An fMRI Study of Emotional Constraints on Partisan Political Judgment in the 2004 U.S. Presidential Election. *Journal of Cognitive Neuroscience*, 18(11):1947–1958, 11 2006. ISSN 0898-929X. doi: 10.1162/jocn.2006.18.11.1947. URL https://doi.org/10.1162/jocn.2006.18.11.1947.

[196] G. Wheeler. Bounded Rationality. In E. N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2024 edition, 2024.

[197] E. Whitaker, E. Trewhitt, M. Holtsinger, C. Hale, E. Veinott, C. Argenta, and R. Catrambone. The effectiveness of intelligent tutoring on training in a video game. *2013 IEEE International Games Innovation Conference (IGIC)*, pages 267–274, 2013. doi: 10.1109/IGIC.2013.6659157.

[198] S. Wijenayake and J. G. and. A review of online social conformity: Outcomes and determinants. *International Journal of Human–Computer Interaction*, 0(0):1–30, 2024. doi: 10.1080/10447318.2024.2424385. URL https://doi.org/10.1080/10447318.2024.2424385.

[199] T. D. Wilson and N. Brekke. Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological bulletin*, 116(1):117, 1994.

[200] M. Wischnewski, N. Krämer, and E. Müller. Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581197. URL https://doi.org/10.1145/3544548.3581197.

[201] M. Yang, H. Arai, N. Yamashita, and Y. Baba. Fair machine guidance to enhance fair decision making in biased people. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642627. URL https://doi.org/10.1145/3613904.3642627.

[202] M. P. Zanna and J. Cooper. Dissonance and the pill: an attribution approach to studying the arousal properties of dissonance. *Journal of personality and social psychology*, 29(5):703, 1974. doi: 10.1037/h0036651. URL https://doi.org/10.1037/h0036651.

[203] R. Zhang, H. Li, H. Meng, J. Zhan, H. Gan, and Y.-C. Lee. The dark side of ai companionship: A taxonomy of harmful algorithmic behaviors in human-ai relationships. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713429. URL https://doi.org/10.1145/3706598.3713429.

[204] Q. Zhu, L. Y.-H. Lo, M. Xia, Z. Chen, and X. Ma. Bias-aware design for informed decisions: Raising awareness of self-selection bias in user ratings and reviews. *Proc. ACM Hum.-Comput. Interact.*, 6 (CSCW2), nov 2022. doi: 10.1145/3555597. URL https://doi.org/10.1145/3555597.