Workshop on Understanding and Mitigating Cognitive Biases in Human-AI Collaboration

Nattapat Boonprakong
University of Melbourne
Melbourne, Victoria, Australia

Gaole He
Delft University of Technology
Delft, The Netherlands

Ujwal Gadiraju
Delft University of Technology
Delft, The Netherlands

Niels van Berkel
Aalborg University
Aalborg, Denmark

Danding Wang
Chinese Academy of Sciences
Beijing, China

Si Chen
University of Illinois at Urbana-Champaign
Champaign, IL, USA

Jiqun Liu
University of Oklahama
Norman, OK, USA

Benjamin Tag
Monash University
Clayton, Victoria, Australia

Jorge Goncalves
University of Melbourne
Melbourne, Victoria, Australia

Tilman Dingler
University of Melbourne
Melbourne, Victoria, Australia

ABSTRACT

AI systems are increasingly incorporated into human decision-making. Yet, human decision-makers are often affected by their cognitive biases. In critical settings, such as medical diagnosis, criminal judgment, or information consumption, these cognitive biases hinder optimal decision outcomes, thereby resulting in dangerous decisions and negative societal impact. The use of AI systems can amplify and exacerbate cognitive biases in their users. In this workshop, we seek to foster discussions on ongoing research around cognitive biases in human-AI collaboration and identify future research directions to understand, quantify, and mitigate the effects of cognitive biases. We will explore cognitive biases appearing in various contexts of human-AI collaboration: what can cause them?; how can we measure, model, mitigate, and manage cognitive biases?; and how can we utilise cognitive biases for the greater good? We will reflect on workshop discussions to form a research community around cognitive biases and bias-aware systems.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Collaborative and social computing**.

KEYWORDS

Cognitive Bias, Human-AI Collaboration, Debiasing

CSCW '23 Companion, October 14–18, 2023, Minneapolis, MN, USA

ACM ISBN 979-8-4007-0129-0/23/10.
https://doi.org/10.1145/3584931.3611284

1 INTRODUCTION

In recent years, Artificial Intelligence (AI) has been increasingly outperforming humans in many tasks, such as classification and forecasting [18, 38]. We have seen a rapid uptake in the deployment of AI systems to complement and support human decision-makers in critical domains: judges use algorithmic risk assessment to determine criminal sentences, doctors rely on machine learning models to diagnose patients, and online media platforms adopt recommendation systems to present users with relevant content items. However, the literature indicates that human decision-makers are not always rational [45]; their decisions are often affected by *cognitive biases* – defined by Tversky and Kahneman [50] as mental shortcuts or heuristics to make faster but less deliberate decisions. Cognitive biases distort our thinking in a way we are often unaware of and can negatively influence decision outcomes. For example, confirmation bias can affect how users interpret and seek information online [1], anchoring bias can induce unfair juridical decisions when presented with multiple pieces of evidence [20], and the Dunning-Kruger effect can hinder appropriate reliance on AI systems [21].

Research has suggested that AI systems can trigger and even amplify cognitive biases in their users [1, 3, 7, 33, 35]. Personalised recommendation systems, for example, optimise content recommendations around the users' preferences and cater predominantly to what users prefer. As a result, such systems risk reinforcing confirmation bias and the echo chamber effect [1, 6, 26]. Moreover, studies have shown that AI explanations can exacerbate our cognitive biases and compromise AI-assisted decision-making, such as trust in AI [33], reliance on AI [5, 10, 21, 39], and interpretation of AI results [27, 46]. Meanwhile, cognitive biases can shape the quality of ground-truth data and thereby influence downstream applications and human evaluations of systems [16, 24], and also influence the outcomes of AI systems [1, 4]. Recommendation systems pick up not only user preferences but also their confirmation bias through their selective information consumption behaviour. As a result, these systems deliver content that, in turn, amplifies users' cognitive biases [1]. A recent example of ChatGPT also demonstrates that it exhibits many biases humans possess, for instance, framing bias and overconfidence bias [12]. With AI systems and cognitive biases forming an interplay that influences human decision-making, it is, therefore, crucial to understand how cognitive biases manifest themselves and how their effects can be mitigated.

A growing body of work has explored how to mitigate cognitive biases in human-AI collaboration. By leveraging AI systems and carefully designing them as a decision aid, Kliegr et al. [27] reviewed twenty cognitive biases in interpreting rule-based machine learning models and associated debiasing techniques. Wang et al. [51] proposed a framework for building explainable AI systems that reduce common cognitive biases, e.g., availability bias and confirmation bias. In addition, some research has proposed technological interventions to counter cognitive biases in human-AI collaboration. Buçinca et al. [10] employed cognitive forcing tools to reduce overreliance on AI. Rastogi et al. [43] introduced a time allocation strategy to mitigate anchoring bias. Bach et al. [3] further evaluated bias mitigation techniques as integrated into the user interface of a clinical decision support tool and identified concerns around the impact on efficiency.

Spanning from literature in behavioural economics, researchers and practitioners have adopted *nudges* as an intervention to counter the undesired effects of cognitive biases [9, 11, 31, 36, 37]. Nudges alter the environment, i.e., the user interface, and subsequently trigger cognitive biases that shift users towards a particular decision or behaviour [48]. Therefore, such methods not only leverage cognitive biases but can also be used to combat their negative effects. For instance, Rieger et al. [44] employed targeted obfuscation on search results to nudge people towards decreasing interaction with attitude-confirming information. By obfuscating content items that may confirm one's beliefs, this nudge taps into the status-quo bias – a tendency to go along with the path of least resistance – and, in turn, helps reduce confirmation bias as users avoid interacting with the obfuscated items.

Nonetheless, effectively mitigating cognitive biases is a challenging task, particularly because of the inherent property of cognitive biases that some individuals are less or more susceptible to biased judgments due to interaction contexts (e.g., domain knowledge, cognitive load, or topic involvement) [8, 30, 35] and individual characteristics (e.g., short-term memory span) [35, 42]. Moreover, different individuals, such as those with varying levels of expertise, possess different mental models of interacting with and understanding AI systems [25, 53]. As a result, not every debiasing intervention would always be effective for every user [21, 35, 44]. Additionally, multiple cognitive biases can manifest at the same time, producing mixed effects that can be difficult to observe and mitigate [2].

The role of cognitive biases in human-AI collaboration has become a growing discourse in the CSCW community. Through different domains, such as HCI, CSCW, Information Retrieval, and Behavioural Economics, diverse forms of studies address the question of how cognitive biases manifest themselves and how they could be effectively mitigated. Therefore, it is important to bridge together insights from different disciplines and create a common ground for future cognitive bias research.

In this workshop, we aim to bring together researchers, practitioners, and designers to jointly seek a better understanding of cognitive biases and solutions to mitigate problems arising from biases. Recent workshops such as *Workshop on Detection and Design for Cognitive Biases in People and Computing Systems* [13] at CHI 2020, *Workshop on Technologies to Support Critical Thinking in an Age of Misinformation* [14] at CHI 2021, and *the Dagstuhl Seminar on Technologies to Support Critical Thinking in an Age of Misinformation* [15] have explored related topics with particular focus on online information consumption and misinformation. This proposed workshop at CSCW 2023 will focus on cognitive biases in the context of human-AI collaboration, where AI systems act as supporting tools for human decision-makers.

2 WORKSHOP GOALS AND THEMES

We aim to foster a discussion about ongoing work on cognitive biases in HCI, provide a common platform to revisit the current research, and establish a research agenda for understanding, quantifying, mitigating, and utilising cognitive biases. Ultimately, we seek to form a research community that works towards the design of *Bias-Aware Systems* [8, 34], defined as computing systems that take into account the cognitive biases of their users. Through creating this community, we aim to build a shared understanding of cognitive biases and methods to measure, utilise, and mitigate their effects. We hope that discussions in this workshop lead to fruitful collaborations that leverage our understanding of cognitive biases in users.

In this workshop, we call for participants to share their research ideas, questions, and opinions with regard to the following themes:

- **Discovering and Identifying Cognitive Biases** We would like to explore mechanisms and components of AI systems that amplify or trigger cognitive biases in their users. In what human-AI collaboration scenarios are cognitive biases involved? Recent research has explored a diverse set of cognitive biases people follow when interacting with explainable AI systems [7, 17, 27].
- **Modelling and Quantifying Cognitive Biases**: An important step towards bias mitigation is to model cognitive biases and measure their extent [4, 40]. However, since users are often unaware of their cognitive biases, it is challenging to know whether cognitive biases are manifesting themselves. Recent research has proposed mathematical frameworks to model cognitive biases [23, 35, 43]. Moreover, some works have explored methods to reliably quantify cognitive biases in-situ using a variety of physiological sensors [8, 19].
- **Novel Approaches to Mitigate Cognitive Biases**: We would like to explore novel methods to mitigate the negative effects of cognitive biases in human-AI collaboration. Existing approaches include *nudging*, i.e., changing the choice environment [11], *boosting*, i.e., fostering metacognitive skills in people [28], and designing *decision support systems* that help users make effective and accurate decisions [51]. We seek to discuss the shortcomings and limitations of existing debiasing approaches and develop future directions.
- **Application Scenarios of Cognitive Biases**: While it is known that cognitive biases negatively affect human decision-making, we would like to explore the use of cognitive biases for the greater

good. Can we imagine scenarios in which cognitive biases actually benefit human-AI collaboration? [29, 32]
- **Impact of the Bias Mitigation**: We seek to explore how bias identification and mitigation strategies can positively and negatively impact AI systems and their users. What benefits do people get if their biases are mitigated? Do we really need to eliminate biases? Is there an alternative way to support human decision-making? Recent research has shown that some debiasing interventions like nudges can harm user autonomy [3, 36] or slow down the interaction [41].
- **Case Studies of Cognitive Biases in Human-AI Collaboration**: Presentation of concrete cases where the prevalence, mitigation, and utilisation of cognitive biases in human-AI collaboration have been investigated.

3 CALL FOR PARTICIPATION

We would like to welcome 20-30 participants for this workshop (excluding the organisers). Participants will be required to contribute a brief statement of interest to the workshop. We accept several forms of submission, including (1) a short research summary or position paper (2-4 pages excluding references) discussing one or more workshop themes or (2) a one-page essay stating motivations for attending this workshop with a short bio. Each submission will be reviewed by the workshop organisers and accepted based on the quality of the submission and the diversity of perspectives to allow fruitful discussions between researchers from different domains, including but not limited to HCI, CSCW, AI, and cognitive psychology. Upon acceptance, we will encourage participants to record a short 3-5 minute video presenting the content of their submission, which will be available to watch before the workshop. We will advertise our workshop and the call for papers through mailing lists, social media, and forums.

4 WORKSHOP SCHEDULE

We propose a one-day workshop with hybrid participation: there will be an option to participate physically at CSCW 2023 and virtually to ensure maximum inclusion. We plan to organise the workshop with the following activities:

- **Introduction** (1 hour): We will welcome participants to this workshop and provide an outline of planned activities, goals, and themes. We will also include a quick ice-breaking activity for participants to get to know each other.
- **Lighting Talks** (1.5 hours): Participants will share their paper submissions. We plan on allocating time for selected papers, and the presentations will be organised under workshop themes. Each presenter will have three minutes to talk about their work and two minutes for Q&A. We aim for this session to be an opportunity for authors to introduce their research and gain feedback from the audience.
- **Two-round Action Group Activities** (2 hours): We will divide participants into *action groups* where each group's theme will associate with concrete scenarios from the submitted position papers and existing pre-workshop discussions, for example, recommender systems, explainable AI, or generative AI. Participants can join the group with the theme they are most interested in. The number of action groups will be determined prior to the workshop. In two rounds, participants will engage in the following activities:
  - **Brainstorming** (40 minutes): Participants will be assigned a brainstorming task and discuss solutions in their action group. Each brainstorming task will be associated with one of the abovementioned themes. In both rounds, there will be at least one organiser facilitating discussion in each group. We plan to source brainstorming tasks from workshop themes: *what are cognitive biases and their causes-triggers-effects* (discovering bias); *what are the measures of cognitive biases* (quantifying bias); *develop interventions to debias* (mitigating bias); *identify application scenarios where cognitive biases are exploited* (utilising bias); and *identify positive and negative impacts from bias mitigation* (impact of bias mitigation).
  - **Knowledge Synthesis** (20 minutes): All participants will reconvene to share what they discussed in their action groups, including key ideas, challenges, and opportunities.
- **Closing Remarks** (30 minutes): We will synthesise key takeaways from discussions and identify the next steps for building a research community on cognitive bias. We will also facilitate follow-up conversations after formally concluding the workshop.

We will incorporate breaks between sessions and social activities into the final schedule.

5 HYBRID SETUP

We plan to utilise the following tools to support and accommodate our hybrid setup:

- **Workshop Website.** We will make the workshop information publicly available on the workshop website[1], including the workshop proposal, call for participation, accepted submissions, workshop program, participant information, and other relevant material.
- **Slack Workspace.** We will set up a dedicated Slack workspace to enable asynchronous communication among workshop participants before, during, and after the workshop. Prior to the workshop, we will also share accepted paper submissions and (if applicable) short videos on the Slack workspace. The organisers will actively monitor discussions on the channels to keep our participants engaged.
- **Zoom Video Conferencing.** We will broadcast in-person workshop presentations, activities, and discussions on Zoom to allow virtual participants to take part in our workshop. The same tool will also allow live closed-captioning to support accessibility.
- **Padlet and Miro Boards.** We will use Padlet and Miro boards to allow participants, both in-person and virtual, to share and note down ideas throughout the workshop. We will encourage workshop participants to take notes on these online sharing tools as they will be accessible for remote participants and future references.

6 WORKSHOP OUTCOMES

The following are the expected outcomes of this workshop.

[1] http://www.critical-media.org/cscw23

- **Forming a research community.** By bringing together researchers and practitioners from different disciplines, we expect to see an exchange of knowledge and future collaborations on research around cognitive biases and bias-aware systems. We will also keep the slack workspace open to workshop participants to continue to develop a community of cognitive bias researchers.
- **Compilation of Cognitive Biases in HCI.** Based on workshop discussions, we intend to document a list of cognitive biases in various HCI contexts and the associated quantification, utilisation, and mitigation strategies, accompanied by empirical studies that explored such biases.
- **Sharing Insights.** We will document the results of the action groups and discussions and make the collected information available to workshop participants and the broader community through an online repository and a public website.

7 ORGANISERS

The organising team of this workshop consists of researchers and experts working in and across CSCW, HCI, AI, and Information Retrieval.

- **Nattapat Boonprakong** is a PhD candidate at the School of Computing and Information Systems, the University of Melbourne. He is interested in Cognition-aware Systems and Empathic Computing. His PhD research is particularly focused on the detection and mitigation of cognitive biases in online information consumption where people often face different opinions.
- **Gaole He** is a PhD candidate at Web Information Systems group of the Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS/EWI), Delft University of Technology. His research focuses on human-centered explainable AI, human-AI decision-making, and knowledge graph reasoning.
- **Ujwal Gadiraju** is an Assistant Professor at the Software Technology Department of the EEMCS faculty at Delft University of Technology. He is a Co-director of the TU Delft Design@Scale AI Lab. His research focuses on *Human-Centered AI and Crowd Computing* to create novel methods, interfaces, systems, and tools capable of overcoming existing challenges at the intersection of HCI and AI in our pursuit to build better AI systems and facilitate better reliance of humans on AI systems. He has (co)-organized workshops and symposiums focused on biases in human computation, crowdsourcing, and human-AI interactions.
- **Niels van Berkel** is an Associate Professor at the Department of Computer Science at Aalborg University. His research focuses on collaboration with real-world intelligent systems, with a focus on overcoming challenges in decision-making. He has (co-)organised various workshops, including on the topic of Human-AI interaction, explainability, and fairness [52].
- **Danding Wang** is an Assistant Researcher at the Institute of Computing Technology, Chinese Academy of Sciences. Her research focuses on human-centered explainable AI, AIGC detection, media forensics, and misinformation detection. She proposed an explainable AI framework that facilitates human reasoning and mitigates cognitive biases [51].
- **Si Chen** is a PhD candidate at the School of Information Sciences at University of Illinois at Urbana-Champaign. She conducts research on human consciousness and metacognition while using AI-driven intelligent learning systems. Additionally, she explores how to ensure inclusivity in such systems for learners with diverse abilities, such as those who are blind and visually impaired.
- **Jiqun Liu** is an Assistant Professor of Data Science and Affiliated Assistant Professor of Psychology at the University of Oklahoma (OU). His research focuses on the intersection of human-centered data science, interactive information seeking/retrieval, and cognitive psychology, and seeks to apply the knowledge learned about people interacting with information in adaptive recommendation, user education and intelligent nudging.
- **Benjamin Tag** is a Lecturer in the Human-Centred Computing Group at Monash University. He researches Human-AI Interaction, Digital Emotion Regulation, and Immersive Analytics with a special focus on inferring mental state changes from data collected in the wild. Benjamin co-organized a series of workshops on cognitive biases [13–15].
- **Jorge Goncalves** is an Associate Professor in the School of Computing and Information Systems at the University of Melbourne. He has conducted extensive research on Human Computation and facilitating Human-AI Interaction. He has also served as Workshops Co-Chair for CHI'19 and CHI'20, and co-organised many successful workshops at leading HCI venues such as CHI, CSCW and Ubicomp [22, 47, 49].
- **Tilman Dingler** is a Senior Lecturer in the School of Computing and Information Systems at the University of Melbourne. He investigates the notion of cognition-aware systems and builds technologies that support users' information-processing capabilities. Tilman instigated a series of recent workshops related to critical thinking and the role of cognitive biases at prime venues, such as CHI and in the prestigious Dagstuhl seminar series [13–15].

REFERENCES

[1] 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

[2] Leif Azzopardi. 2021. Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (Canberra ACT, Australia) *(CHIIR '21)*. Association for Computing Machinery, New York, NY, USA, 27–37. https://doi.org/10.1145/3406522.3446023

[3] 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. 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 16, 14 pages. https://doi.org/10.1145/3544548.3581513

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

[5] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. 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 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[6] 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

[7] 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

[8] 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

[9] 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

[10] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (apr 2021), 21 pages. https://doi.org/10.1145/3449287

[11] 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

[12] Yang Chen, Meena Andiappan, Tracy Jenkin, and Anton Ovchinnikov. 2023. A Manager and an AI Walk into a Bar: Does ChatGPT Make Biased Decisions Like We Do? *Available at SSRN 4380365* (2023). https://dx.doi.org/10.2139/ssrn.4380365

[13] 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

[14] 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

[15] 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

[16] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. 2021. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 9. 48–59.

[17] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael J. Muller, and Mark O. Riedl. 2021. The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations. *CoRR* abs/2107.13509 (2021). arXiv:2107.13509 https://arxiv.org/abs/2107.13509

[18] Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. 2018. When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research* 62 (2018), 729–754.

[19] Christopher G. Harris. 2019. Detecting Cognitive Bias in a Relevance Assessment Task Using an Eye Tracker. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications* (Denver, Colorado) *(ETRA '19)*. Association for Computing Machinery, New York, NY, USA, Article 36, 5 pages. https://doi.org/10.1145/3314111.3319824

[20] Yugo Hayashi and Kosuke Wakabayashi. 2017. Can AI Become Reliable Source to Support Human Decision Making in a Court Scene?. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17 Companion)*. Association for Computing Machinery, New York, NY, USA, 195–198. https://doi.org/10.1145/3022198.3026338

[21] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. 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 113, 18 pages. https://doi.org/10.1145/3544548.3581025

[22] Danula Hettiachchi, Mark Sanderson, Jorge Goncalves, Simo Hosio, Gabriella Kazai, Matthew Lease, Mike Schaekermann, and Emine Yilmaz. 2021. Investigating and Mitigating Biases in Crowdsourced Data. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing* (Virtual Event, USA) *(CSCW '21)*. Association for Computing Machinery, New York, NY, USA, 331–334. https://doi.org/10.1145/3462204.3481729

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

[24] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12. https://doi.org/10.1145/3290605.3300637

[25] Kori Inkpen, Shreya Chappidi, Keri Mallari, Besmira Nushi, Divya Ramesh, Pietro Michelucci, Vani Mandava, LibuŠe Hannah VepŘek, and Gabrielle Quinn. 2023. Advancing Human-AI Complementarity: The Impact of User Expertise and Algorithmic Tuning on Joint Decision Making. *ACM Trans. Comput.-Hum. Interact.* (mar 2023). https://doi.org/10.1145/3534561 Just Accepted.

[26] 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

[27] 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

[28] 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.

[29] Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. 2011. Mining Behavioral Economics to Design Persuasive Technology for Healthy Choices. 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, 325–334. https://doi.org/10.1145/1978942.1978989

[30] 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

[31] 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

[32] Q. Vera Liao and Kush R. Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *CoRR* abs/2110.10790 (2021). arXiv:2110.10790 https://arxiv.org/abs/2110.10790

[33] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-Distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 408 (oct 2021), 45 pages. https://doi.org/10.1145/3479552

[34] Jiqun Liu. 2023. *A Behavioral Economics Approach to Interactive Information Retrieval: Understanding and Supporting Boundedly Rational Users*. Vol. 48. Springer Nature.

[35] 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

[36] 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

[37] 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. 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 759, 19 pages. https://doi.org/10.1145/3544548.3581058

[38] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. 2020. International evaluation of an AI system for breast cancer screening. *Nature* 577, 7788 (2020), 89–94. https://doi.org/10.1038/s41586-019-1799-6

[39] 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. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 340–350. https://doi.org/10.1145/3397481.3450639

[40] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 3 (2020), e1356.

[41] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 102 (nov 2019), 15 pages. https://doi.org/10.1145/3359204

[42] Gordon Pennycook and David G. Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated

reasoning. *Cognition* 188 (2019), 39–50. https://doi.org/10.1016/j.cognition.2018.06.011 The Cognitive Science of Political Thought.

[43] 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, Article 83 (apr 2022), 22 pages. https://doi.org/10.1145/3512930

[44] 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

[45] 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.

[46] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, Textual or Hybrid: The Effect of User Expertise on Different Explanations. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 109–119. https://doi.org/10.1145/3397481.3450662

[47] Benjamin Tag, Sarah Webber, Greg Wadley, Vanessa Bartlett, Jorge Goncalves, Peter Koval, Petr Slovak, Wally Smith, Tom Hollenstein, Anna L Cox, and Vassilis Kostakos. 2021. Making Sense of Emotion-Sensing: Workshop on Quantifying Human Emotions. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers* (Virtual, USA) *(UbiComp '21)*. Association for Computing Machinery, New York, NY, USA, 226–229. https://doi.org/10.1145/3460418.3479272

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

[49] Garreth W. Tigwell, Zhanna Sarsenbayeva, Benjamin M. Gorman, David R. Flatla, Jorge Goncalves, Yeliz Yesilada, and Jacob O. Wobbrock. 2019. Addressing the Challenges of Situationally-Induced Impairments and Disabilities in Mobile Interaction. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3290607.3299029

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

[51] 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

[52] Sofia Yfantidou, Dimitris Spathis, Marios Constantinides, Tong Xia, and Niels van Berkel. 2023. FairComp: Workshop on Fairness and Robustness in Machine Learning for Ubiquitous Computing. In *Adjunct Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing* (UbiComp'23 Adj.). to appear.

[53] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. 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 437, 21 pages. https://doi.org/10.1145/3544548.3581388