Report on the CHIIR 2024, SIGIR 2024, and SIGIR-AP 2024 tutorials on Characterizing, Evaluating, and Mitigating Cognitive Biases in Interactive and Generative IR

Jiqun Liu
The University of Oklahoma
Norman, OK, USA
jiqunliu@ou.edu

Leif Azzopardi
University of Strathclyde
Glasgow, Scotland
leifos@acm.org

Abstract

This event report presents an overview of a series of tutorials conducted at CHIIR 2024, SIGIR 2024, and SIGIR-AP 2024, all centered around the emerging, socially significant theme of *cognitive biases* in Interactive Information Retrieval (IIR) and Generative IR (GenIR). Each tutorial focused on a critical aspect of this growing research area, offering attendees a combination of theoretical foundations from psychology and behavioral economics, empirical studies, hands-on methodological discussions, and forward-looking design exercises in IIR and GenIR applications. As IR systems increasingly integrate generative and conversational interfaces, understanding the interplay between human cognitive biases and AI-driven information environments becomes not only essential for advancing theory but also vital for designing responsible and user-centered search and chat systems. These tutorials provided participants with analytical tools, practical frameworks, and collaborative opportunities to rethink ranking and evaluation, user modeling, and system design through the lens of cognitive psychology and behavioral economics. The strong engagement across all three SIGIR venues signals a growing momentum toward bias-aware, human-centric IR research.

Date: March–December 2025.

Website: https://beiir.github.io/.

1 Introduction

Recent advancements in interactive search and recommendation systems, especially those powered by large language models (LLMs), have transformed how people seek, process, and evaluate information [Zhu et al., 2023; Acharya et al., 2023; Li et al., 2023]. At the same time, the field of Interactive Information Retrieval (IIR) has recognized the need to go beyond purely rational models to incorporate the *boundedly rational*, *heuristic-driven* nature of users [Liu, 2023a; Azzopardi, 2021]. Cognitive biases — systematic deviations from rational judgment — are now seen as not

just peripheral quirks but key components of search behavior with significant implications for IR evaluation, algorithmic fairness, user trust and AI audit [Chen et al., 2024; Liu, 2023b].

To foster community-wide dialogue, knowledge sharing, and potential collaboration within this research space, we organized a three-part tutorial series on cognitive biases in IR across CHIIR 2024 (Sheffield, UK) [Azzopardi and Liu, 2024b], SIGIR 2024 (Washington, DC) [Liu and Azzopardi, 2024], and SIGIR-AP 2024 (Tokyo, Japan) [Azzopardi and Liu, 2024a], covering a variety of research focuses, approaches and perspectives, and regional IR communities. Each tutorial was designed to build upon the previous one, evolving from theory-driven analysis to hands-on methodological guidance and IR evaluation strategies, and finally, to system design within the context of Generative IIR. Collectively, these tutorials aimed to 1) raise awareness of human biases in search, evaluation, and GenIR interactions, 2) offer tools and methods to identify, test, and mitigate human biases in IIR sessions and batch evaluations, and 3) equip participants and students to design more psychologically realistic and responsible user models, IR systems, and evaluation measures in both traditional and generative search contexts.

2 Overview of Tutorials

This section offers an overview for each of the tutorials, which address different research questions and aspects of cognitive bias research in Interactive IR and GenIR contexts.

2.1 ACM CHIIR 2024 Tutorial

The first tutorial "*Search under Uncertainty: Cognitive Biases and Heuristics – Tutorial on Modeling Search Interaction using Behavioral Economics*", delivered at ACM CHIIR 2024 in Sheffield (UK) as a half-day event, provided a theoretical and empirical introduction to cognitive biases and mental shortcuts in search under uncertainty. We began by contextualizing these biases within classic dual-process theories of reasoning and judgment, particularly Tversky and Kahneman [1974]'s *System 1* (fast, intuitive) and *System 2* (slow, deliberative) thinking. Participants engaged in a *cognitive bias survey* [1] — "*How biased are you*?"—that used validated psychological experiments to illustrate and test common biases, such as anchoring, confirmation bias, expectation disconfirmation, decoy effect, and framing effect [Kahneman, 2003; Thaler, 2016; Tversky and Kahneman, 1992]. The survey includes a set of typical questions and hypothetical decision scenarios used in cognitive psychology and behavioral economics, and is included to facilitate participants' discussions and also make the tutorial more engaging. We presented the collective results from the survey to participants before sharing theories and empirical findings on cognitive biases from behavioral experiments and research on information seeking and retrieval. This opening survey segment established a cognitive foundation for understanding how such tendencies shape users' intents, search behavior, document judgment, and search experience.

Guided by the tutorial organizers' survey studies on cognitive biases and bounded rationality in search (e.g., [Azzopardi, 2021; Liu, 2023a]), the remainder of the tutorial introduced empirical studies that examined these biases in concrete information-seeking and IR contexts (e.g., [Draws et al., 2021; White, 2013; Rieger et al., 2021; Agosto, 2002; Liu and Han, 2020]). For example,

[1] Our cognitive bias survey is accessible here: https://forms.gle/ApwbfFYBQqVBWQTm7

we discussed how the order of search results can influence trust and judgment, how exposure to biased content can reinforce prior beliefs and expectations, how biases like the decoy effect and framing effect can distort user preferences and utility judgment during evaluation on document batches and search engine result pages (SERPs), and how the behavioral impact of cognitive biases may vary across search tasks, search stages, and user intents of varying types. We also reviewed emerging frameworks for mapping the "space" of cognitive biases and heuristics across different online information search stages, such as query and prompt formulation, browsing, document judgment, and post-search reflection, and across different domains, from medical decision-making, trip and event planning, to political information seeking and assessment. This initial tutorial clarifies the implicit connection between the knowledge on human biases and heuristics and the challenges in human-centered IR research, and highlight the importance of understanding and assessing the impact of biases on users' document judgment and search experiences. Participants left the tutorial with a clearer sense of how these biases challenge traditional assumptions in user modeling and system evaluation, and how the knowledge about bounded rationality can inform the design of more realistic formal models, useful search systems, and offline IR evaluation resources.

2.2 ACM SIGIR 2024 Tutorial

Built upon the theories and empirical findings introduced at the CHIIR 2024 tutorial above, the SIGIR 2024 tutorial "*Search under Uncertainty: Cognitive Biases and Heuristics – A Tutorial on Testing, Mitigating and Accounting for Cognitive Biases in Search Experiments*", held in Washington, D.C., shifted focus toward methodology and evaluation, with an emphasis on how to test for and mitigate cognitive biases in IR research and applications. Structured as a full-day event, the tutorial introduced the theoretical foundations and research progresses on identifying and assessing human biases in IR and related fields (e.g. information seeking, recommender systems, HCI), and guided participants through the design of behavioral and system evaluation experiments that isolate and examine specific biases in search sessions, document judgments, and conversational interactions. We explored how to operationalize human bias in IR contexts, define appropriate dependent variables, measures and labeling procedure, and ensure internal validity when working with user studies. A substantial portion of the session focused on how biases can confound user judgments, IR system evaluation (in both online and offline metrics) and task performance, and how to account for these issues in experimental and observational settings. For example, when searchers concerned about vaccine safety first encounter a snippet echoing their prior beliefs, confirmation bias nudges them to click similar results and artificially boost engagement metrics. In offline relevance judging, showing assessors a high-scoring "gold" answer at the start often anchors subsequent grades around that reference, warping measures like nDCG. Also, adding an obviously inferior third subscription tier may trigger the decoy effect and steer both users and ranking algorithms toward the mid-priced option and skewing conversion-rate evaluations. Participants discussed several key questions related to study design, such as: 1) how to design search tasks, bias variables, and experimental conditions to examine the "clean" effect of certain biases? 2) how should we characterize and measure the potential positive effects of biases and heuristics (e.g. improving the efficiency in routine task completion; reducing cognitive loads); 3) how to identify the "main bias" in specific search scenarios and differentiate the impacts of different cognitive

biases within the same search session; and 4) how to incorporate human bias features into offline evaluation metrics and assess the risk of potential bias triggers on SERPs.

Going beyond bias detection, this tutorial also delved into practical strategies for system-driven in-situ bias mitigation (e.g., [Wang and Liu, 2024a; Zerveas et al., 2022; Wu et al., 2023]), including user interface interventions (e.g. biased result obfuscation and warning labels), bias-aware ranking adjustments, and search interface nudging techniques. These de-biasing strategies could be applied and evaluated at different aspects and levels, such as reminding users of potential biases, directly reducing biased contents, and adjusting ranking algorithms based on users' individual levels of *vulnerabilities* to certain biases [Chen et al., 2025; Liu et al., 2025]. The follow-up group discussions allowed participants to critique real-world interface layouts and experimental designs, propose improvements at varying levels, and reflect on the ethical considerations of studying and intervening in users' cognitive processes and limits in search interactions. The group discussion also touches on the challenges in evaluating and mitigating cognitive biases in *generative search interactions*, such as behavioral signals needed for bias prediction, replicable human bias labeling on generated responses, impact on user engagement and perceived cognitive loads, and conversational bias mitigation. Many attendees expressed interest in continuing this work and discussion through future tutorials, collaborative studies and workshops, shared datasets and open-access tools, and more psychologically grounded evaluation experiments and reusable resources.

2.3 ACM SIGIR-AP 2024 Tutorial

At the SIGIR-Asia Pacific 2024 in Tokyo (Japan), the full-day tutorial entitled "*Evaluating Cognitive Biases in Conversational and Generative IIR*" tackled the emerging challenges and opportunities in bias evaluation and mitigation posed by Conversational and Generative Information Retrieval (GenIR). As conversational agents become more prevalent in search interfaces, their interaction with human cognitive biases grows more complex, and potentially more consequential in varying judgment and decision-making scenarios. In addition, LLMs as retriever and information generator could inherit and amplify human biases in their responses and automated judgments, which may reinforce existing beliefs and biases of users [Bender et al., 2021]. This tutorial explored the ways in which GenIR systems could not only reflect users' biases but also actively amplify them through persuasive responses, tailored system explanations, and emotionally resonant language.

We reviewed and discussed recent studies that investigate how conversational systems and LLM-enabled chatbots may reinforce confirmation bias and threshold priming biases, subtly steer user judgments and preferences through framing and decoy effects, or affect users' confidence and decision certainty. In addition, during the tutorial presentation session, we explained how the behavioral patterns of users and LLMs in bias-triggering scenarios echo the findings from previous behavioral experiments and IR evaluation research on cognitive biases (e.g., [Liu and Han, 2020; Eickhoff, 2018; Draws et al., 2021; Rieger et al., 2024]). This session highlights how the knowledge about human judgments and actions under biases and heuristics can inform the modeling and evaluation of IR evaluation. Moving on to the hands-on participatory session, a key feature of the tutorial was an interactive "co-design" mini workshop in which participants worked in small groups to design user experiments and system functionalities targeting these emerging risks. Each group proposed or was assigned a subtopic, such as anchoring in LLM suggestions or reference-dependence in AI explanations, and tasked with developing a study design

that could reveal how these biases manifest and measure bias impacts in generative interfaces. To facilitate design activity and discussion, we proposed a list of *template questions* and leave the specific biases and variables open to the participants. Example questions include: 1) Are people susceptible to a certain bias when interacting with (y)? Under what conditions, are people more susceptible?; 2) Does one bias have a greater impact or effect than another?; 3) How can we mitigate bias (z) when using (y) to improve (x)?; 4) How can we test whether [Application y] is capitalizing /manipulating given people's [Cognitive Bias z]. Participants were asked to pick a cognitive bias or a research paper on cognitive bias and consider how they might design an example to explore or test the impact of the bias. Participants shared their design insights and peer review comments in a collaborative review session, leading to vibrant discussion around experimental challenges, evaluation measures, ethical boundaries, and methodological innovation. The tutorial concluded with a brief discussion on future research opportunities and challenges in the field, such as identifying signals of biased behaviors for predicative modeling, understanding and assessing boundedly rational users' interactions with conversational systems, differentiating negative impacts of biases and heuristics from positive experiences (e.g. improved search efficiency and task performance, reduced cognitive loads) in real-time interactions, and developing reliable benchmarks for evaluating system performance in debiasing information interactions.

3 Reflections and Cross-Tutorial Insights

The tutorials collectively revealed an urgent need to reconsider how cognitive assumptions are built into IR systems and research methods. One key takeaway across the three events was that biases are not just anomalies in individual judgment; they are deeply intertwined with the structures, interfaces, and incentives that shape interactive IR interaction of varying modalities [Liu, 2023a]. From the presentation of ranked lists to the framing of LLM prompts and the design of evaluation metrics, every layer of the IR pipeline can serve as a trigger, or a buffer, for cognitive bias.

A major point of consensus was the mismatch between the ideal user assumed in formal IR modeling and the *boundedly rational* user observed in real-world behavior [Liu, 2023a]. While traditional models often assume that users carefully assess options, weigh relevance criteria, and make optimal selections, empirical studies repeatedly show that users rely on mental shortcuts. These include rank position bias (favoring top-ranked results), confirmation bias (seeking information that aligns with prior beliefs), and anchoring (fixating on the first or most salient piece of information) [Wang and Liu, 2024b; White, 2013; Eickhoff, 2018; Ji et al., 2024]. These tendencies are not trivial: they can skew evaluations of system performance, lead to inaccurate document judgments, and reinforce disinformation or low-quality content in online information seeking.

Participants also recognized that cognitive biases are not static. They are context-sensitive and shaped by the task at hand, the user's prior experience, search results and interface affordances. For example, the same user might be more susceptible to confirmation bias when searching for emotionally charged health information than when researching technical topics. Biases can also interact with each other, and users may simultaneously experience anchoring and framing effects during a single session, which complicates both detection and intervention. These layered dynamics highlight the need for more nuanced approaches to studying IR challenges, including mixed-methods designs that blend behavioral logging, surveys, and experimental manipulation.

From an evaluation standpoint, the tutorials surfaced growing concerns with current metrics that rely on implicit user behavior as a proxy for relevance and usefulness. Click-through rates, dwell time, and session length may all be shaped by biases rather than by system performance and SERP quality. This complicates claims of usefulness judgment, user satisfaction and effectiveness, particularly when interfaces themselves may nudge or manipulate search behavior. There was enthusiasm around integrating reference-dependent models [Chen et al., 2023], as well as bias-aware metrics that account for decision framing, loss aversion, anchoring and decoy effects, and information overload [Chen et al., 2025; Liu and He, 2025]. Several participants noted that measuring interaction experience and user perception under the impact of biases and heuristics, not just retrieval outcome, should be a central goal of IIR evaluation moving forward.

Perhaps the most animated discussions occurred in the SIGIR-AP tutorial, where participants debated the ethical implications of bias in conversational and GenIR. Questions emerged such as: What responsibilities do designers of LLM interfaces have in shaping belief formation? Should systems warn users when their queries or interactions appear biased or unfair? Can LLMs be used constructively to counteract cognitive distortions and manipulations? Many acknowledged the dual role of GenAI as both amplifier and potential mitigator of bias. For instance, while conversational systems can reinforce confirmation bias by adapting to users' beliefs, they also offer new features, such as multi-perspective generation and explicit uncertainty expression, which may encourage critical thinking. These tensions call for ethical and policy exploration.

Beyond individual learning, the tutorials fostered a sense of collective momentum. Across CHIIR, SIGIR, and SIGIR-AP, attendees came from diverse subfields , including IR, information behavior, HCI, psychology, library studies, computer and data science, and AI ethics, yet shared a common desire to rethink how IR systems interact with and affect human cognition. The group discussions, particularly the collaborative design exercises at SIGIR-AP, were a fertile ground for new research questions and design insights, many of which can inspire follow-up interdisciplinary projects, students' theses, and external grant proposals. Attendees expressed interest in organizing dedicated workshops, submitting co-authored position papers and reports, conducting experiments to assess bias impacts and evaluate mitigation techniques, and developing shared evaluation infrastructure to advance this agenda. We believe that research on characterizing, evaluating, and mitigating human biases can bring together the insights from both human-centered IR studies and system-oriented simulation and evaluation experiments, and facilitate the dialogue among IR researchers employing different approaches in studying users and evaluating systems [Liu, 2023a].

4 Future Directions

The growing interest in human cognitive bias opens up several promising trajectories for research, system evaluation, infrastructure development, and community building. Drawing from participant feedback across the tutorials, we identify five key directions for future work:

4.1 Development of Shared Resources and Toolkits

Aligned with the evaluation focus of IR community, a recurring theme across the three tutorials was the pressing need for concrete, shareable tools that enable researchers and developers to assess cognitive biases in IR systems. Building on participants' contributions, we propose establishing

a publicly accessible "bias-aware IR evaluation toolkit". Its core modules would comprise: 1) validated bias-assessment instruments (short-form cognitive-reflection tests, in-situ user expectation measures, belief-strength and anchoring-susceptibility scales) packaged as language-agnostic surveys; 2) bias-aware logging templates and tools for popular IR testbeds that capture query reformulations, click latency, ranking snapshots, and other bias-relevant behavioral signals; and 3) a living repository of interface design patterns that have been shown to reveal or mitigate biases in search, recommendation, and chat scenarios. Each module should come with concise step-by-step protocols, covering lab studies, large-scale online experiments, and field deployments.

To keep the toolkit actionable and current, we recommend a lightweight governance model: A small steering group seeds the repository with gold-standard data, then issues quarterly "pull-request sprints" inviting the community to contribute new probes, logging hooks, and mitigation patterns. A complementary leaderboard can track A/B test results for common tasks (e.g., health queries, polarizing news topics) using bias-aware evaluation metrics, while a "failure log" documents negative or null findings to prevent redundant effort. With clear contribution guidelines, ready-to-run examples, and periodic virtual workshops, the toolkit would lower the entry barrier for newcomers, facilitate replication research, and accelerate cumulative progress toward bias-resilient IR systems.

4.2 Toward Bias-Aware Evaluation Metrics and Benchmarks

Current effectiveness measures, such as click-through rate, dwell time, topical relevance scores and relevance-based offline metrics [Chen et al., 2017; Moffat et al., 2012], usually assume that searchers evaluate every result with perfect rationality and quantifiable utility maximization goals. However, positional anchoring, reference dependence, and decoy effects weave themselves into the interaction trace, warping both behavior logs and offline labels and obscuring a system's true contribution to user utility. A next-generation framework must embed cognitive parameters into its formalism and adjust gains and losses relative to a user's shifting reference points, discounting engagements that follow anchoring patterns, and reporting confidence bands that capture individual uncertainty.

Several converging threads can move this vision from principle to practice. Open-source simulators that inject controlled bias profiles into synthetic user agents would let researchers stress-test candidate metrics before running expensive live studies. In tandem, benchmark corpora linking raw interaction traces with ground truth debiasing labels, such as fairness in result clicking and evaluation, post-task confidence ratings, belief strength surveys or heuristic-use eye tracking markers, would allow head-to-head comparisons of calibration techniques such as propensity reweighting and reference-dependent normalization. A community leaderboard could then report both traditional effectiveness scores and bias-adjusted variants (e.g. *DEJA-VU* score [Chen et al., 2025]), signaling to system designers that relevance and cognitive resilience are co-equal optimization targets. Through combining simulation, richly annotated datasets and transparent comparisons, bias-aware evaluation can become a routine checkpoint in the deployment of IR systems.

4.3 Experimental Methodologies for Conversational and GenIR

As GenAI continues to reshape the IR landscape, research must evolve to match the new challenges it introduces. Studying cognitive bias in GenIR settings calls for novel experimental designs

and system interventions that go beyond static search result sets and better capture the variations in users' interactions with generative systems. These include interactive protocols for measuring framing effects in dialog, longitudinal tracking of user confidence, belief change and expectation variations, and experimental manipulations that vary LLM tone, verbosity, or explanation format [Liu et al., 2025]. Controlled studies with human participants interacting with LLMs under varied cognitive load and emotional states could yield insights into not just whether bias occurs, but why it arises in these dynamic settings and how we could mitigate them. In addition, researchers should also explore how LLMs and AI-enabled agents may exhibit human-like cognitive biases in automated judgments and decisions (e.g. misinformation assessment, hiring and candidate review, legal consultation), and how these biases can be identified and proactively addressed in human-AI collaborative work [Lyu et al., 2024; Zhou et al., 2024; Chen et al., 2024].

Building on these foundations, future work should seek to integrate mixed-methods approaches, such as combining fine-grained interaction logs with neuro-physiological measures (e.g., eye movement features, galvanic skin response, EEG-based brain signals [Mostafa and Gwizdka, 2016; Ye et al., 2024; Jacucci et al., 2019; Boonprakong et al., 2024; Edwards and Kelly, 2017]), to construct a richer picture of how bias unfolds over the course of a conversation. Embedding adaptive debiasing mechanisms into the dialogue loop, such as on-the-fly perspective balancing, uncertainty-aware prompt recommendations, or "devil's advocate" generation modes, would allow us to test not only the efficacy of these interventions but also their impact on user trust, perceived autonomy, cognitive load, and task efficiency. Moreover, simulation environments powered by synthetic user agents parameterized with known bias profiles could facilitate rapid, large-scale stress-testing of GenIR systems before deployment in naturalistic settings. Finally, understanding the role of human biases and heuristics is essential for developing more realistic and useful support for human-AI interaction and AI-assisted decision making [Gajos and Mamykina, 2022]. As applications move into high-stakes domains, it is imperative to develop open-source benchmarking suites that pair bias-stress tests (counterfactual prompting, adversarial framing) with longitudinal field studies, ensuring that mitigation strategies are not only empirically grounded but also ethically sound.

4.4 Design and Evaluation of Bias-Mitigation Strategies

Once robust bias detectors are in place, researchers can also explore nudging techniques and interventions that actively blunt harmful bias while safeguarding user autonomy and trust. Mitigation levers may operate at three levels. Interface nudges can foreground under-represented viewpoints, flag potentially manipulative framings, or introduce brief reflection prompts that slow "System-1" click habits. Adaptive system behaviors, such as dynamic perspective balancing, confidence weighted re-ranking, or probabilistic result shuffling, may offer deeper control when surface-level nudges prove insufficient. At the most granular tier, personalized debiasing adjusts these strategies to an individual's measured susceptibility profile, tuning the breadth of viewpoint diversification or the strength of cautionary messaging in real time. Striking the right balance is crucial: mitigation must be potent enough to offset bias, transparent enough to command user confidence, and friction-light enough to preserve task flow and decision-making processes.

Generative IR systems and LLM-powered chat interfaces add fresh opportunities and risks. Because these models can synthesize multiple perspectives on demand, they have the potential to counter confirmation and anchoring bias through cognitive forcing functions, such as present-

ing contrasting rationales, annotating responses with uncertainty estimates, or prompting users to articulate counter-arguments and evidences before accepting an answer. Mitigation modules for GenIR might inject diversity constraints or debiasing rewards into beam search, or rely on retrieval-augmented generation (RAG) [Lewis et al., 2020; Gao et al., 2023] that surfaces evidence clusters from ideologically distinct sources. Evaluation pipelines should therefore pair offline bias-stress-tests (e.g., counter-factual prompting, synthetic user simulations) with longitudinal field studies that track belief and expectation revision, decision quality, and perceived transparency and cognitive load, especially in domains where the stakes are high, such as health, finance, hiring, and legal research. Embedding these LLM-specific mitigation into open-source libraries, along with pre-registered study templates and ethics checklists, would enable rapid iteration while ensuring that bias reduction remains empirically grounded and ethically defensible.

4.5 Community Infrastructure and Collaboration

Given the interdisciplinary nature of this agenda, coordination and communication are vital. We plan to expand the tutorial website into a more comprehensive platform for sharing literature, experimental templates, code, and event announcements. We are also planning to organize additional tutorials, Webinars, and/or regular workshop series under SIGIR or CHIIR that focuses on cognition-aware IR and evaluation, and build connections to similar workshops and collaborative events under other related communities, such as SIGCHI (e.g., [Boonprakong et al., 2023; Dingler et al., 2020]), ACM FAccT, and ASIS&T. Such a venue would allow researchers working on bias, trust, user modeling, human factors, and GenAI interaction to connect, collaborate, and shape a cohesive and sustainable research program. In the longer term, we envision partnerships between academia, industry, and civil society organizations to address the real-world consequences of biased IR systems and interactions. As generative systems increasingly mediate access to health, legal, and financial information in online spaces, designing for *cognitive resilience* — and not just convenience or efficiency — will be a defining challenge for the IR field and broader, rapidly growing GenAI research and applications.

5 Conclusion

By integrating the knowledge about cognitive biases with interactive and generative IR research, the tutorial series guided participants from foundational concepts through empirical research in information seeking and IR contexts, bias-aware evaluation metrics for retrieved documents and chat system outputs, and design of future bias identification and mitigation techniques in generative IR scenarios. Participant feedback underscored strong enthusiasm for incorporating human bias parameters into user models and system evaluations, shaping future SIGIR workshops, informing doctoral research agendas and funding proposals, and enriching IR curricula and training programs. In mapping clear pathways for benchmark development, practical system integrations, and educational initiatives, this work lays a durable foundation for advancing cognitive resilience in IR and offers compelling opportunities for both established researchers and emerging scholars.

Acknowledgments

The first author was partially supported by a seed grant from the Data Institute for Societal Challenges at the University of Oklahoma and a faculty research award from Microsoft. The second author was partially supported by the DoSSIER project under European Union's Horizon 2020 research and innovation program, Marie Sklodowska-Curie grant agreement No. 860721.

References

Arkadeep Acharya, Brijraj Singh, and Naoyuki Onoe. Llm based generation of item-description for recommendation system. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1204–1207, 2023.

Denise E Agosto. Bounded rationality and satisficing in young people's web-based decision making. *Journal of the American society for Information Science and Technology*, 53(1):16–27, 2002.

Leif Azzopardi. Cognitive biases in search: a review and reflection of cognitive biases in information retrieval. In *Proceedings of the 2021 ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 27–37, 2021.

Leif Azzopardi and Jiqun Liu. Evaluating cognitive biases in conversational and generative iir: A tutorial. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 287–290, 2024a.

Leif Azzopardi and Jiqun Liu. Search under uncertainty: Cognitive biases and heuristics-tutorial on modeling search interaction using behavioral economics. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 427–430, 2024b.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

Nattapat Boonprakong, Gaole He, Ujwal Gadiraju, Niels Van Berkel, Danding Wang, Si Chen, Jiqun Liu, Benjamin Tag, Jorge Goncalves, and Tilman Dingler. Workshop on understanding and mitigating cognitive biases in human-ai collaboration. In *Companion Publication of the 2023 ACM SIGCHI Conference on Computer Supported Cooperative Work and Social Computing*, pages 512–517, 2023.

Nattapat Boonprakong, Kaixin Ji, Ziyi Ye, Benjamin Tag, Damiano Spina, Tuukka Ruotsalo, and Flora D Salim. Advancing physiological methods for human-information interaction. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 976–979, 2024.

Nuo Chen, Jiqun Liu, and Tetsuya Sakai. A reference-dependent model for web search evaluation: Understanding and measuring the experience of boundedly rational users. In *Proceedings of the ACM web conference 2023*, pages 3396–3405, 2023.

Nuo Chen, Jiqun Liu, Xiaoyu Dong, Qijiong Liu, Tetsuya Sakai, and Xiao-Ming Wu. Ai can be cognitively biased: An exploratory study on threshold priming in llm-based batch relevance assessment. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 54–63, 2024.

Nuo Chen, Jiqun Liu, Hanpei Fang, Yuankai Luo, Tetsuya Sakai, and Xiao-Ming Wu. Decoy effect in search interaction: Understanding user behavior and measuring system vulnerability. *ACM Transactions on Information Systems*, 43(2):1–58, 2025.

Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. Meta-evaluation of online and offline web search evaluation metrics. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 15–24, 2017.

Tilman Dingler, Benjamin Tag, Evangelos Karapanos, Koichi Kise, and Andreas 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*, pages 1–6, 2020.

Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. 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*, pages 295–305, 2021.

Ashlee Edwards and Diane Kelly. Engaged or frustrated? disambiguating emotional state in search. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 125–134, 2017.

Carsten Eickhoff. Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 162–170, 2018.

Krzysztof Z Gajos and Lena Mamykina. Do people engage cognitively with ai? impact of ai assistance on incidental learning. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, pages 794–806, 2022.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2, 2023.

Giulio Jacucci, Oswald Barral, Pedram Daee, Markus Wenzel, Baris Serim, Tuukka Ruotsalo, Patrik Pluchino, Jonathan Freeman, Luciano Gamberini, Samuel Kaski, et al. Integrating neurophysiologic relevance feedback in intent modeling for information retrieval. *Journal of the Association for Information Science and Technology*, 70(9):917–930, 2019.

Kaixin Ji, Sachin Pathiyan Cherumanal, Johanne R Trippas, Danula Hettiachchi, Flora D Salim, Falk Scholer, and Damiano Spina. Towards detecting and mitigating cognitive bias in spoken conversational search. In *Adjunct Proceedings of the 26th International Conference on Mobile Human-Computer Interaction*, pages 1–10, 2024.

Daniel Kahneman. Maps of bounded rationality: Psychology for behavioral economics. *American economic review*, 93(5):1449–1475, 2003.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Lei Li, Yongfeng Zhang, Dugang Liu, and Li Chen. Large language models for generative recommendation: A survey and visionary discussions. *arXiv preprint arXiv:2309.01157*, 2023.

Jiqun Liu. A behavioral economics approach to interactive information retrieval. *The Information Retrieval Series*, 48, 2023a.

Jiqun Liu. Toward a two-sided fairness framework in search and recommendation. In *Proceedings of the 2023 ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 236–246, 2023b.

Jiqun Liu and Leif Azzopardi. Search under uncertainty: Cognitive biases and heuristics: a tutorial on testing, mitigating and accounting for cognitive biases in search experiments. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3013–3016, 2024.

Jiqun Liu and Fangyuan Han. Investigating reference dependence effects on user search interaction and satisfaction: A behavioral economics perspective. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 1141–1150, 2020.

Jiqun Liu and Jiangen He. Boundedly rational searchers interacting with medical misinformation: Characterizing context-dependent decoy effects on credibility and usefulness evaluation in sessions. In *Proceedings of the 2025 ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 154–167, 2025.

Jiqun Liu, Jamshed Karimnazarov, and Ryen W White. Trapped by expectations: Functional fixedness in llm-enabled chat search. *arXiv preprint arXiv:2504.02074*, 2025.

Yougang Lyu, Xiaoyu Zhang, Zhaochun Ren, and Maarten de Rijke. Cognitive biases in large language models for news recommendation. *arXiv preprint arXiv:2410.02897*, 2024.

Alistair Moffat, Falk Scholer, and Paul Thomas. Models and metrics: Ir evaluation as a user process. In *proceedings of the seventeenth Australasian document computing symposium*, pages 47–54, 2012.

Javed Mostafa and Jacek Gwizdka. Deepening the role of the user: Neuro-physiological evidence as a basis for studying and improving search. In *Proceedings of the 2016 acm on conference on human information interaction and retrieval*, pages 63–70, 2016.

Alisa Rieger, Tim Draws, Mariët Theune, and Nava 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*, pages 189–199, 2021.

Alisa Rieger, Tim Draws, Mariët Theune, and Nava Tintarev. Nudges to mitigate confirmation bias during web search on debated topics: Support vs. manipulation. *ACM Transactions on the Web*, 18(2):1–27, 2024.

Richard H Thaler. Behavioral economics: Past, present, and future. *American economic review*, 106(7):1577–1600, 2016.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131, 1974.

Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5:297–323, 1992.

Ben Wang and Jiqun Liu. 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*, pages 227–237, 2024a.

Ben Wang and Jiqun Liu. Understanding users' dynamic perceptions of search gain and cost in sessions: An expectation confirmation model. *Journal of the Association for Information Science and Technology*, 75(9):937–956, 2024b.

Ryen White. Beliefs and biases in web search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12, 2013.

Zhangyi Wu, Tim Draws, Federico Cau, Francesco Barile, Alisa Rieger, and Nava Tintarev. Explaining search result stances to opinionated people. In *World Conference on Explainable Artificial Intelligence*, pages 573–596. Springer, 2023.

Ziyi Ye, Qingyao Ai, and Yiqun Liu. Brain-computer interface meets information retrieval: Perspective on next-generation information system. In *Proceedings of the 1st International Workshop on Brain-Computer Interfaces (BCI) for Multimedia Understanding*, pages 61–65, 2024.

George Zerveas, Navid Rekabsaz, Daniel Cohen, and Carsten Eickhoff. Mitigating bias in search results through contextual document reranking and neutrality regularization. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2532–2538, 2022.

Hanyu Zhou, Sreya Muppalla, Zeynep Yildiz, Miryung Kim, Amit Sahai, Nanyun Peng, et al. Measuring psychological depth in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17162–17196, 2024.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*, 2023.