Chapter 7
Implications and New Directions for IR Research and Practices

Abstract Previous chapters have thoroughly discussed recent advances and progresses in IR formal user models and behavioral economics research on human bounded rationality in decision-making. As presented in Chap. 6, some recent studies in IR, information seeking, and recommendation have empirically confirmed the impacts of human biases and heuristics on users' search interactions, judgments of information items, and reactions to personalized recommendations and partially incorporate the knowledge of bounded rationality into developing user behavior prediction models, system evaluation metrics, and bias-aware re-ranking algorithms. Taking a step forward from previous discussions, this chapter will introduce existing unresolved research gaps and open challenges from bounded rationality perspective and discuss the main research questions, practical implications, and new directions of our behavioral economics approach for various sub-areas of IR studies.

7.1 Background

Previous chapters have thoroughly discussed recent advances and progresses in IR formal user models and behavioral economics research on human bounded rationality in decision-making (with empirical evidences associated with both behavioral patterns and neural correlates). Contrasting the specific assumptions, model setups, and findings from these two areas of research clarify a series of gaps between simulated rational agents and real-world users engaging in search interactions under varying tasks. These gaps motivated us to reflect on the existing oversimplified assumptions and rational user models and encouraged us to explore ways in which we could extend the assumptions about user characteristics and behavioral patterns and also enhance existing formal models. As presented in Chap. 6, some recent studies in IR, information seeking, and recommendation have empirically confirmed the impacts of human biases and heuristics on users' search interactions, judgments of information items, and reactions to personalized recommendations and partially incorporate the knowledge of bounded rationality in developing user behavior prediction models, system evaluation metrics, and bias-aware re-ranking algorithms. Taking a step forward from previous discussions,

J. Liu, *A Behavioral Economics Approach to Interactive Information Retrieval*, The Information Retrieval Series 48, https://doi.org/10.1007/978-3-031-23229-9_7

Chap. 7 will introduce existing unresolved research gaps and open challenges from bounded rationality perspective and discuss the main research questions, practical implications, and new directions of our behavioral economics approach for various sub-areas of IR studies. We hope that our synthesis of the insights from related areas and new bias-aware research agenda can be of help for students and researchers who are interested in further investigating more specific human-bias-related IR problems and leveraging the learned knowledge in enhancing intelligent search systems.

7.2 Characterizing Bounded Rationality in IR

When making decisions under uncertainty, people are often boundedly rational due to a series of individual characteristics and situational limits, such as cognitive and perceptual biases, mental shortcuts, as well as limited resources and support (Simon, 1955; Kahneman, 2003). In the context of IR, previous studies have explored a set of the widely examined human cognitive biases and described their implicit connections to users' information search behaviors, document judgment thresholds, and whole-session evaluations (e.g., Azzopardi, 2021; Eickhoff, 2018; Liu & Han, 2020; Scholer et al., 2013). To further enhance our understanding of boundedly rational search decisions, researchers may need to address several general limitations.

First, it is worth noting that people's boundedly rational decisions and judgments usually involve perceived *multidimensional changes*, *gains*, and *losses* (Kahneman, 2003). Although existing IR research has examined several types of human biases and heuristics (see Chap. 6), many of them only focused on one or two dimensions associated with the impacts of biases. One of the widely examined dimensions is *relevance judgment* as a large body of user studies, and offline IR evaluation experiments include external relevance labeling as part of the standard experimental setup. However, users' biased perceptions and decisions could occur in other dimensions as well, such as the judgments of document credibility and usefulness, acceptance of different types of search recommendations, as well as the experience with certain search interfaces. For instance, the impact of decoy results and threshold priming could not only be triggered by the relevance labels of documents examined in sequence but also initiated by the difference in document presentation (e.g., text only or augmented with relevant images; presented as regular organic search results or vertical results) and perceived document credibility. In addition, users' search satisficing strategy and aspiration level may also be multidimensional in nature and are influenced by the perceived gains and losses on a variety of facets of search interactions, rather than depending on query-document relevance (*qrel*) only. More broadly, from the reference-dependence perspective (e.g., Tversky & Kahneman, 1991), users' pre-search and in situ preferences and expectations may also involve different dimensions, such as search interactions and costs, system effectiveness, and document quality, as well as overall search experience. These different dimensions of references could also change over time as a search session proceeds and may have different weights in users' search decision-making and whole-session remembered

utility. It is difficult to characterize diverse references and associated temporal variations (e.g., changes of SERP quality across different queries; changes of users' preferences over diverse subtopics) with only one or two ground-truth labels.

Next, related to the first limitation, when characterizing the temporal changes along different dimensions and associated with varying human biases, researchers also need to explore the *interactions among different dimensions of search interactions and biases*. For instance, based on e-commerce click and purchase logs, Ge et al. (2020) identified the mutual reinforcements between individual users' interests and the biases in item exposure in recommender systems, which confirmed the echo chamber effect in personalized product recommendations. Azzopardi (2021) also discussed possible compounding effects caused by two or more cognitive biases on searchers. For example, individuals' decisions are often heavily influenced by the initial information available in a given sequence (*Primacy effect*; Jones et al., 1968). This primacy effect may couple with *anchoring bias*: when a user evaluates results in a SERP, the first item presented or examined may be considered as most relevant and used as an anchoring point for judging the relevance and credibility of following documents. Also, for each individual search decisions in a session, such as query reformulation, document clicking, and search stopping, the user may be influenced by both in situ reference points (e.g., search results with varying types of framing) and pre-search existing beliefs and preferences (*Confirmation bias*, Nickerson, 1998; White, 2013).

Although mutual reinforcement effect could occur among diverse human biases (Azzopardi, 2021), different biases may also compete with each other for people's attention and relatively higher weights in final decision-making. For instance, suppose a user has pre-search doubts about the effectiveness of a certain brand of vaccine. During the information search process, the user may be actively searching for results that confirm or is aligned with their pre-search beliefs. However, when the top ranked search results contradict with the pre-search expectations, the confirmation bias might be mitigated by the anchoring bias or in situ reference dependence as the user may consider the initially encountered or top ranked search results as most relevant. This impact of in situ reference may be weaker if the disconfirming results are ranked in lower positions on the SERP. Therefore, to comprehensively investigate the interaction between confirmation bias and anchoring bias (as well as the interplay of other human biases), researchers may also need to take into consideration the roles of several IR-specific factors, such as search query features, search result presentations (e.g., as organic search results or vertical blocks), and adaptive learning to rank algorithms.

With respect to the behavioral impacts of bounded rationality, researchers from multiple disciplines (including IR) have extensively studied the negative impact on search performance, document judgment quality, and overall experiences. As a result, a series of negative effects and biased decision strategies have been identified through the behavioral experiments where researchers start with assumed negative

effect of human biases.[1] However, the potential *positive effects* of human biases remain understudied. For instance, cognitive biases and heuristics may reduce the complexity of decision-making processes (Azzopardi, 2021). Relying on a set of simple rules and mental shortcuts, individuals may be able to quickly obtain satisficing or good-enough outcomes without processing a large amount of new information (Kahneman, 2011). Also, people may be more likely to be affected by cognitive biases when facing conflicting information (which usually increases the uncertainty in option evaluation and decision-making). Thus, using certain mental shortcuts may help reduce the uncertainty and improve the efficiency in decision-making activities. This could be of high value to users, especially in scenarios where timeliness is more important than optimized accuracy. Future IR researchers should actively explore the positive effects of human biases and heuristics in making search decisions, judging information items, and performing information-intensive work tasks. More broadly, estimating the positive impact of bounded rationality may also enable researchers and system engineers to build more comprehensive computational models of real-time human decisions under bounded cognitive resources (Gershman et al., 2015).

Apart from the research gaps above, characterizing human biases and boundedly rational decisions in IR is also *methodologically challenging*. To capture the "pure effect" of human biases and heuristics, behavioral economics researchers often choose to observe human decisions within well-controlled, simplified, and sometimes unrealistic experimental settings (e.g., Kahneman, 2003; Thaler, 2016; Weber & Camerer, 2006), such as gambling with two options, selling or buying one item, or deciding the treatment plan with the complete knowledge of the probability of cure associated with each alternative. These simplified experimental settings allow researchers to extract one decision-making segment out of complex real-world settings and directly observe the phenomenon of bounded rationality with other contextual variables (e.g., work task characteristics, other people's opinions and actions, domain knowledge, and information seeking skills) being controlled. However, interactive search sessions often tend to be complex, dynamic, and involve contextual factors of multiple levels, such as action level, query level, search task level, as well as motivating task level. Even in controlled user study settings where users conduct search activities under a predefined search task, it is still difficult to redesign or deconstruct the complex search processes into a set of single-decision-based simplified experiments. In addition, to address the problem of observing and modeling the interactions between diverse human biases, researchers may not be able to restrict the decision-making experiment within an oversimplified setting.

As shown in Fig. 7.1, within a simplified representation of general methodological spectrum, researchers need to find the *scientifically reasonable* and *practically accessible* balance between the two directions or sides: On the one side, (over)-simplified task and study designs that are widely applied in behavioral economics

[1] This phenomenon may also confirm the existence of confirmation bias in IR research on bounded rationality.

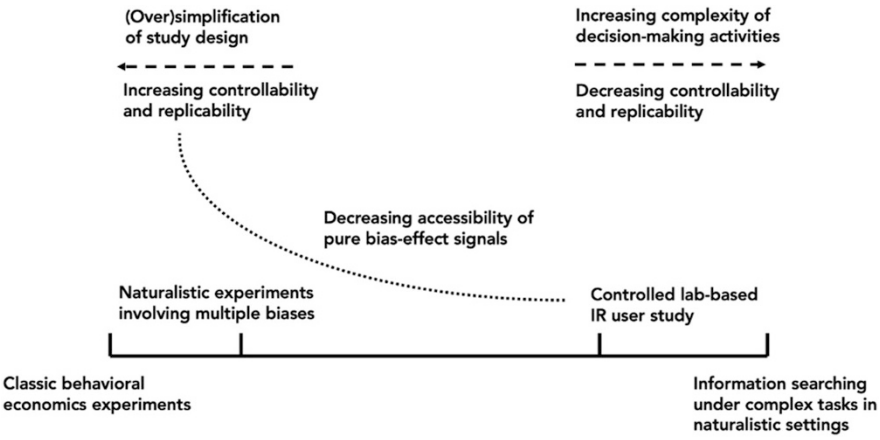

Fig. 7.1 Methodological challenges in IR research on bounded rationality

and cognitive psychology experiments can increase the chance of observing pure behavioral effects caused by human biases and heuristics. The well-controlled experimental settings can make the bias trigger for salient (e.g., a clear decoy option among different alternatives) and thus more likely to generate significant, testable behavioral variations. On the other side, however, we also need to consider the level of *realism* of the simulated tasks and decision-making settings. Although the simplified environments can better facilitate the investigation on human biases, it may affect the generalizability and practical value of the research findings. The behavioral effect measured in simple experiments may not be practically meaningful in complex IR settings where multilevel search decisions are mixed with each other. Also, it is worth noting that several types of biases are identified under artificially constructed experimental sessions; this may be because the study participants are not intrinsically motivated to find out the real credible information items (Azzopardi, 2021). For instance, a high school student participant may not have a clear motivation for learning about available retirement plans, except for the compensation payment for their participation. As a result, the participant may be more likely to stop searching at satisficing results, which present clear but overestimated signals of bounded rationality to the researchers who may be subject to *observer-expectancy effect* (Rosenthal, 1976). This issue of participants' motivations in task completion is a common issue in crowdsourcing studies (e.g., Law et al., 2016; Posch et al., 2019; Rogstadius et al., 2011) and may cause extra risk for user studies on human biases and heuristics. However, when completely departing from controlled lab experiment contexts and customized interventions, researchers may also find it difficult to identify and access reliable signals that indicate boundedly rational actions and capture the implicit deviation of biased decisions from mathematically optimal outcomes.

Crowdsourcing-based user experiments, especially the ones focusing on the judgment and labeling of information items (e.g., Eickhoff, 2018; Maddalena

et al., 2016; Roitero et al., 2022), enable researchers to partially characterize users' evaluation decisions under the influence of cognitive biases that emerge from individuals' naturalistic settings. However, how to go beyond a single slice of search process and reasonably approximate whole-session search interaction experience that involves multistage decision-making still remains an open challenge. To address this challenge, researchers will need to both design effective tasks, interfaces, and interventions that can be naturally implemented in real-life settings and also identify new measures and signals for capturing the multidimensional effects of human biases, heuristics, and situational limits and depicting boundedly rational decisions.

This section summarizes the existing limitations and research gaps in terms of characterizing bounded rationality in IR. For each limitation, we have identified specific research problems to be addressed and suggested possible paths for future studies. The knowledge learned through exploring and characterizing users' bounded rationality can provide a solid behavioral and psychological basis for designing new search and ranking algorithms, user interface components, and recommendations, as well as bias-aware evaluation metrics.

7.3 Development of Bias-Aware Interactive Search Systems

The second part of our research agenda adopting the behavioral economics perspective focuses on the open challenges we need to address regarding the development of bias-aware interactive search systems. Our ultimate goal is that the bias-aware search algorithms and systems can take into account the impacts of both algorithmic biases and human biases and proactively address the potential negative effects from users' biased perceptions, judgments, and search decisions, especially in complex search tasks of varying types.

Taking a step forward from the discussions in the above section, researchers need to properly present diverse types of human biases, heuristics, as well as other situational factors that contribute to boundedly rational decisions and estimate corresponding parameters in updated formal user models with real-world search interaction data. As summarized in Chaps. 4 and 6, there are a large body of behavioral economics experiments and IR user studies that described and statistically tested the effects of various human biases on judgment and decision-making (Azzopardi, 2021; Kahneman, 2003). However, the knowledge of bounded rationality accumulated in a variety of disciplines has rarely been incorporated into the design of formal user models. This research gap can be considered as part of the broader, deeper disconnection between information seeking community and IR community: although a variety of online information seeking behavior models have been proposed in diverse specific settings, populations, and task scenarios, many of them have not been introduced or represented in formal, computational models of user behavior in IR experiments, partly because of the descriptive nature and significant individual differences embedded in information seeking models and

practical limitations in available training datasets and ground truth labels (Liu, 2022).

The *enhancement of formal user models* under a bounded rational framework would start with the multifaceted extension of simplified rational assumptions discussed in Chap. 5. For instance, when estimating the attractiveness and probability of examination before clicking, researchers should consider not only the textual features of search result surrogates and the rank position of the document but also the past history of browsing, clicking, and judgments, especially the reference levels, anchoring points, and in situ preferences hidden in past search interactions. Also, due to the *threshold priming* effect (cf. Scholer et al., 2013), users may keep adjusting and calibrating their thresholds of relevance judgments during a sequence of query-driven search iterations. Regarding usefulness judgments, researchers will also need to examine search task facets (e.g., Li & Belkin, 2008; Liu, 2021) and monitor the distance between the current document and the overarching search tasks. Due to the subjective nature and since individual differences involved usefulness judgments, the calibrated thresholds may not regress to a relatively stable value as it is expected in relevance judgments (Thomas et al., 2022). Thus, researchers may need to design and empirically test different forms of customized *task-document distance* measures and see which one(s) best capture the user's in situ usefulness perceptions.

The dynamic nature of references, judgment thresholds, and information need often lead to unexpected deviations of users' examination and clicking behaviors from the predictions of traditional click models. Therefore, incorporating explicit representations of in situ references and implicit judgment thresholds into click models may improve the accuracy in both unbiased relevance estimation and click prediction. The connections between previous references extracted from actions, documents, and explicit feedback (if available) and current search actions could be represented by extra edges in the *session flow* of graph-based click models (Lin et al., 2021). Adopting a data-driven approach, the weights of hidden edges among pre-search and in situ references, implicit thresholds, and current document features could be learned through neural networks from search log data containing both intra-session and inter-session information. In addition to the graph-based method, researchers could also adopt a personalized click model (PCM) approach and incorporates users' reference points and cognitive biases into click models as part of the user factors. For instance, within Shen et al. (2012)'s PCM framework, user biases could be represented as elements of user matrix, which in turn shapes the Gaussian prior of the document attractiveness parameter in click modeling.

In addition to predicting clicking and characterizing within-SERP browsing, researchers have also developed a series of rational models for formally modeling interactive search sessions. One of the common modeling approaches is to deconstruct users' search sessions into the transitions of a fixed set of *phases* or *states*. The phases and states are either defined under a starting theoretical framework and a set of axioms in a top-down fashion (e.g., Dungs & Fuhr, 2017; Zhai, 2016) or empirically extracted from users' search logs and explicit annotations (e.g., Hendahewa & Shah, 2013; Liu et al., 2020; Liu & Yu, 2021). With the state-based

framework, researchers have proposed a variety of optimization algorithms to iteratively maximize the search effectiveness or scores of evaluation metrics (e.g., nDCG, average precision, document usefulness) that measure ranking performances and SERP qualities (Luo et al., 2014; Liu & Shah, 2022; Zhang & Zhai, 2016). To enhance existing state transition models, researchers can incorporate bias-related factors into the models and estimate their impacts on state transitions. For instance, with the same local retrieval outcome, a relative change (perceived as a gain or loss) in the relevance of search result surrogate or page dwell time may significantly affect the probabilities of browsing continuation and transition to the stage of query reformulation. Also, encountering a document that confirms the user's existing beliefs and expectations may result in an unexpectedly high probability of clicking and relevance score that deviate from the average probability estimated based on past search behaviors and the textual features of current documents.

Another possible approach to integrate bounded rationality factors into session modeling is to add the hidden bias-aware states to the framework of observable behavioral states. Specifically, for example, in addition to the explicit transitions among query, search result snippet examination, and clicking, researchers can also characterize and monitor the reference-dependent state. With the knowledge of pre-search beliefs and preferences, researchers can estimate and label the (dis)-confirmation state of each document and also represent the query/SERP-level perceived utility based on the state associated with each document. In formal modeling, researchers can still focus on relevance-based scores as the main component of utility modeling and at the same time add factorized residuals and use latent confirmation-state factors to depict individuals' deviations from the "global model" built upon query-document relevance. Similarly, researchers could also model reference-dependent states with respect to the anchoring bias by investigating the anchoring effects of initially encountered documents, subtopics, and associated opinions from content generators and other users. In addition, based on the studies on user expectations in interactions with search and management information systems (e.g., Lankton & McKnight, 2012; Liu & Shah, 2019; Venkatesh & Goyal, 2010), researchers could estimate users' general multifaceted expectations regarding rewards and costs or efforts in search interactions and predict the *expectation (dis)confirmation* state at different moments of real-time search sessions. Behavioral and textual signals that indicate a negative expectation disconfirmation (i.e., search efficiency or SERP quality lower than pre-search expectations) may serve as useful features for predicting the changes of subsequent search tactics and users' in situ thresholds and criteria for usefulness judgment and search satisfaction.

On the system side, one of the central topics that connect multiple sub-areas of IR research is *learning to rank* (LTR). L2R refers to the research that applies machine learning (ML) techniques in training ranking models based on annotated relevance labels and implicit feedback (e.g., clicks) from users' search logs (Li, 2011). The goal of LTR research is to train a learning function that produces a ranking score $\pi_\mu(d)$ based on the feature vector of each document d so that the ranking result based on $\pi_\mu(d)$ would be the same as the result of ranking by the *intrinsic relevance* of

documents. According to Ai et al. (2021), this ranking optimization goal can be formally written as:

$$\mu^* = arg_\mu \, min \, \mathcal{L}(\mu) = arg_\mu \, min \int_q^Q l\left(\pi_\mu, \boldsymbol{r}_q\right) \, dP(q) \tag{7.1}$$

where $\boldsymbol{r}_q$ refers to the perfect ranking generated based on the ground-truth intrinsic relevance of documents and Q refers to the set of all queries or topics q involved in ranking. $l(\pi_\mu, \boldsymbol{r}_q)$ represents the loss of local ranking computed based upon the ranked list of retrieved documents and their relevance levels. One of the key challenges is LTR experiments to improve the *unbiasedness* in ranking, especially in situations where the implicit feedback (in particular, clicking) is noisy and affected by different types of biases. To achieve unbiased learning to rank (ULTR), many IR researchers have designed and tested multiple click models and formal assumptions, based on features of query-document pairs, rank position information, and sequences of user actions, in order to facilitate the extraction of reliable relevance signals from noisy and biased click logs (e.g., Ai et al., 2018; Craswell et al., 2008; Joachims, 2002). Apart from the research efforts on reducing rank position bias, some recent studies focus on the behavioral side of LTR and have adopted inverse propensity weighting (IPW) (cf. Joachims et al., 2017) in addressing trust bias and recency effects (e.g., Agarwal et al., 2019; Chen et al., 2019; Vardasbi et al., 2021).

Following the line of research introduced above, researchers may be able to make further progress in ULTR research by taking a broader range of human biases into consideration. As discussed in previous chapters, apart from the widely examined rank position bias, knowledge of human biases and heuristics employed in search can enhance our understanding of the motivations behind clicking behavior. For instance, for documents being ranked at similar positions, users' examinations and clicks may be biased toward the documents that confirm their pre-search beliefs or are consistent with their initially encountered information in one or multiple aspects (e.g., subtopic, opinion, sentiment), due to the effects of confirmation bias and anchoring. Also, in addition to the document features that most ranking algorithms focus on, users' probability of clicking on certain documents may be affected by adjacent documents that may be perceived as a decoy option. In click modeling, identifying potential decoy search results along varying dimensions (e.g., relevance and presentation of search result surrogate, perceived credibility of documents) may be included as part of the modeling of local contextual factors. From a broader reference-dependence and CBDT perspective (e.g., Gilboa & Schmeidler, 1995; Tversky & Kahneman, 1991), the probability of clicking on a document may also be biased due to a recent experience of examining and reading similar documents under similar motivating tasks. This similarity could be represented with a vector containing multiple elements, such as the specific contents and involved subtopics, type of information sources, general opinion and sentiment, as well as other salient textual and graphical features. A past bad experience (e.g., a long reading session

without much useful information obtained) accessible in the user's current short-term memory may result in a low estimated utility or predicted loss associated with the current document, which, in turn, leads to a fairly low probability of clicking (i.e., loss aversion bias). When evaluating search results under uncertainty (e.g., exploring an unfamiliar domain), users may tend to avoid potential risks and losses and click the search results that seem to be familiar and involve low risk of search failure to them (i.e., risk aversion bias). Actively debiasing noisy clicks based on the knowledge of human biases, heuristics, and search expectations could be useful for further improving the unbiasedness, generalizability, and reliability of ULTR algorithms.

In addition to predicting users' clicks, modeling search sessions, and improving LTR algorithms, researchers should also explore possible recommendation techniques and methods that can reactively or even proactively address the potential negative effects of bounded rationality. In the context of IR, query auto-completion and suggestion are two common and widely employed forms of search support. To stimulate critical thinking and careful decision-making, interactive IR researchers employed query recommendations to present search terms (e.g., survey, comparison, evidence) that could encourage critical thinking in search evaluation and query reformulation (Yamamoto & Yamamoto, 2018). Based on the results from crowdsourcing-based experiments, researchers found that the *query priming* with critical-thinking terms motivated users to issue more queries and revisited SERPs more frequently. Also, under the query priming condition, users were exposed to more Web pages that encourage evidence-based decision-making. From the bounded rationality perspective, the query priming techniques designed in Yamamoto and Yamamoto (2018) presented a positive anchoring point for users and motivated them to examine and click more search results that are aligned with the critical thinking terms advocated through recommended search terms. Similarly, Ong et al. (2017) manipulated the initial *information scent levels* for participants by changing the number and distribution of relevant documents on the first result page in a session. The results indicate that when improving the number and positioning of relevant results on the first result page, the participant's ability to locate relevant results were also improved in both desktop-based and mobile search environments. Therefore, in addition to passively react to biased implicit feedback (e.g., clicks) and search decisions (e.g., early abandonment of query), search systems could proactively adjust the initial query recommendations and the presentations of SERP items in order to mitigate possible negative effects of mental shortcuts and help users find the desired information items. Beside query priming and relevant documents, researchers can also explore other dimensions of SERP (Speicher et al., 2015), such as informativeness, information density, possible confusions and distractions, as well as the potential scrolling and interaction efforts, in order to estimate users' perceived costs in a more accurate manner (instead of assuming fixed equal costs of each action across all queries and topics) and address biased search decisions with more SERPs of higher levels of usability and accessibility.

Built upon the above discussions on different aspects of IR and bounded rationality, a broader vision we aim to pursue in future research is developing *bias-aware*

intelligent task support (BITS) systems. The ultimate goal of this BITS system is to predict and proactively address the negative effects of *both* algorithmic biases and human biases in real-time information interactions and offer scalable, reliable, and unbiased informational support for users engaging in complex search and motivating tasks of varying types. As introduced and explained throughout this book, achieving the vision of BITS would require the completions of a series of interrelated research tasks, including:

- Reflecting on and redefining the basic unit of user and search session modeling and moving from absolute-outcome-based variables and measures to gain- and loss-based units and measures. A hidden challenge related to this is identifying potential reference points and estimating their weights in real-time search decision-making.
- Leveraging the knowledge of bounded rationality in extending simplified assumptions about users and their rules of decision-making and enhancing formal user models applied in predicting single search actions (e.g., examination of search result surrogates, clicking, and query reformulation), characterizing and simulating whole-session interaction processes, and building reusable evaluation metrics.
- Based on the predicted user behaviors and judgments within the limits of bounded rationality, adaptively adjusting the available tools and components for improving search support, such as query auto-completion and suggestion, learning to rank algorithms, and other usability dimensions of SERPs and overall search interface.
- Building new bias-aware evaluation framework that comprehensively assesses the performance of BITS system, at both single-iteration and whole-session levels, in terms of satisfying users' information needs and mitigating the negative impacts of both algorithmic biases and cognitive biases on users' search interaction, judgment of information items, as well as post-search information-intensive decision-making.

Based on the discussions above, Fig. 7.2 illustrates the basic structure and main components of the envisioned BITS system, which in real-time search sessions is evaluated in terms of both enhancing search effectiveness and addressing the effects of interrelated biases from both human and algorithm sides. In the third part of our research agenda, we will focus on the problem of *bias-aware evaluation* and discuss the ways in which we can leverage the knowledge of bounded rationality in better assessing the support that an intelligent search system offers for users engaging in complex tasks.

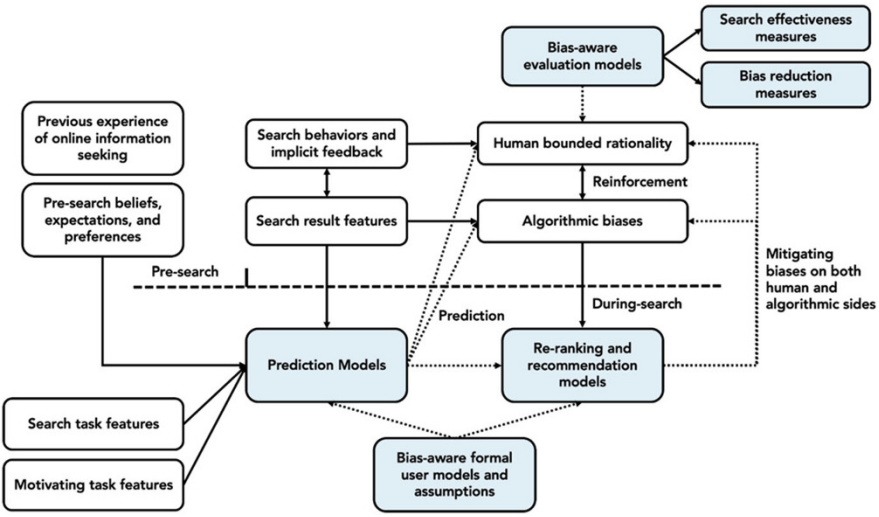

Fig. 7.2 The structure of BITS system. The main activities of the BITS system are denoted by dotted lines

7.4 Bias in Multiple Forms and Modalities of Search Interactions

As people increasingly rely on intelligent information systems for accessing information and making decisions, human biases could occur and operate in various modalities of information search interactions. Recent progress in automatic speech recognition (ASR), deep learning (DL), and natural language processing (NLP) opens new opportunities for research, applications, and technological innovations in *Conversational Information Seeking* (CIS) (Trippas et al., 2018; Yan et al., 2022). Instead of typing queries and studying logical operators in advanced search, users are enabled to simply speak natural language queries and receive visual or verbal respond from IR systems. Furthermore, systems can also help users refine their queries and better express their intentions by asking system-initiated clarifying questions (Sekulić et al., 2021; Zhang et al., 2018). Inspired by multimodal human conversations and interactions, IR researchers have developed models and techniques to go beyond standard SERP presentations, combine multiple channels of information interactions (e.g., spoken/voice-based, text-based, visual information), and facilitate the design, implementation, and evaluation of multimodal CIS (Deldjoo et al., 2021). Under the impact of human biases and *contextual triggers* (e.g., initially encountered information, existing beliefs and expectations, perceived gains and losses), users may change their inputs on multiple channels, such as queries and natural questions, conversational cues and intonations, eye movements, gestures, and facial expressions. Information communicated through these ways can be perceived as indicators of user preferences by search and recommendation

systems and reinforce existing cognitive and algorithmic biases. Differing from traditional desktop searches, conversational search and retrieval systems are often used in in situ factual searches (e.g., looking for nearby convenience store and gas station), little-effort judgments, and intuitive quick decision-making (e.g., selecting an item to purchase among multiple similar items in online shopping sites, picking a restaurant for a quick meal, making subscription decisions for online services). Under these circumstances, people's behaviors are more likely to be affected by cognitive biases, mental shortcuts, as well as the rapid operation of System 1 (Kahneman, 2003). As many HCI researchers seek to build human-like features into conversational systems and treat them as social agents (Thomas et al., 2021), the cognitive and behavioral effect of human biases may occur more naturally and unconsciously in information interactions.

Given the challenges and opportunities associated with human biases in CIS, researchers need to first identify new signals and CIS-specific features that allow them to better identify and estimate the *risks* of human biases generating negative effects at the moment of interactions. The risk estimation will require the knowledge of both user characteristics (e.g., existing beliefs and expectations, prior experiences with search and CIS) and potential contextual triggers of biases (e.g., results that confirm certain misleading beliefs, biased presentation of varying perspectives). Furthermore, researchers need to develop bias-aware user models that predict users' judgments and decisions based on the signals from ongoing interactions and estimated risks of certain biases. The prediction results could be leveraged as part of the basis for developing and implementing adaptive and even proactive recommendations and interventions for effective debiasing. This process could be achieved through modifying the internal result ranking algorithms based on the predicted risks of biases, changing system-initiated questions to reminders of potential biases in mixed initiative CIS systems, and adaptively adjusting online evaluation metrics according to users' search intentions and the nature of predicted biases. Researchers can evaluate the performance of systems and the associated intervention techniques based on the extent to which they can predict and mitigate the risks of biased behaviors and judgments in search interactions.

In addition to CIS, the behavioral impact of user biases could also happen and need to be addressed in other modalities of search interactions, such as mobile search (Lagun et al., 2014; Mao et al., 2018), augmented-reality-based search (Büschel et al., 2018), and tangible IR (Leon et al., 2019; Jansen et al., 2010). Under varying modalities of interactions, researchers will need to identify different sets of signals, rebuild models to predict potential biases, and identify contextual triggers in system outputs and the problems that motivate users to interact with systems. With new signals collected, researchers can infer the perceived informational gains and search costs at different stages and under varying local intentions and utilize them as features in predicting users' search decisions (e.g., query/question reformulation or engaging with current responses, accepting or skipping system recommendations) and in situ experiences (e.g., search satisfaction, perceived cognitive load, overall level of engagement). Knowledge learned about human biases in multimodal

information searching will allow researchers to better capture and address the impact of biases in real time.

7.5 Bias-Aware Evaluation and FATE in IR

To develop trustable AI-assisted intelligent systems (which include BITS and other types and modalities of intelligent search systems), researchers and system designers need to recognize the impacts of both algorithmic and human biases in AI and understand how and where they contribute to harms (Schwartz et al., 2022). In particular, users who are vulnerable to the negative effects of certain biases (e.g., due to prior beliefs, lack of domain knowledge and algorithmic awareness) should be protected from the recommendations and system interventions that leverage the knowledge about their biases for profits and engagements. Apart from formally modeling users and improving retrieval algorithms, the insights about bounded rationality discussed in previous chapters and sections can also be applied in enhancing multiple aspects of user-oriented IR evaluation. For example, in query-level or single-SERP evaluation, representing and estimating *anchoring effect* would be useful for improving the correlation between (anchoring-aware) evaluation metrics and users' levels of search satisfaction (Chen et al., 2022). Beyond Chen et al. (2022)'s work, in future evaluation experiments, researchers should go beyond relevance labels and explore other dimensions of initially encountered documents that may affect users' perceived utility. Based upon the identified anchoring point and other potential references, researchers can develop reference-aware evaluation metrics that assess the *perceived* performance of search systems based on the estimated search gains and losses.

In addition, at task level, researchers need to study pre-search existing beliefs, expectations, as well as their origins, such as past search experience under similar tasks, existing opinions and stereotypes, as well as other people's actions and opinions. Then, during search sessions, researchers can build and test expectation-aware evaluation metrics that consider both pre-search general expectations and in situ dynamic expectations and examine the impacts of expectation disconfirmation on users' search strategies and effectiveness, post-search decision-making, as well as the overall levels of satisfaction and engagement. Similar to query-level evaluation, the exploration of expectation disconfirmation states will also require researchers to investigate multiple facets of search tasks and system outputs (Liu & Shah, 2019), as different facets and dimensions may have significantly different impacts on users' decisions under uncertainty. Besides, in whole-session retrospective evaluation, researchers need to examine a series of key moments, such as initial experience, peak values, and last or most recent experience (Kahneman, 2003; Liu & Han, 2020; Liu et al., 2019) and examine their respective effects on users' remembered utility obtained from the session. Moving forward from the discrete reference points identified in studies on peak-end rule and recency effects, future research could design and implement a more generalizable and flexible *continuous weight*

distribution that considers the varying impacts of all moments of search interactions. From this perspective, in evaluation analysis, different types of search tasks could be linked to different kinds of reference-aware weight distributions that partially characterize the associated search sessions.

More broadly, the behavioral economics research agenda presented in this chapter also motivates us to reflect on the problem of *fairness*, *accountability*, *transparency*, and *ethics* (FATE) in the context of IR evaluation. Although contemporary IR systems provide rapid ubiquitous access to information, they also encode and even amplify the biases, inequalities, and historical gaps through information presentation and recommendation. Addressing the limitations and inherited bias in algorithms would require a deep understanding of human bias as well. Specifically, for instance, when seeking to improve algorithmic fairness and avoid discrimination against different populations and communities, search systems need to take into consideration the specific thinking style and hidden cognitive biases associated with different task types, work environments, and cultural backgrounds (e.g., Ma-Kellams, 2020). Without proper regulation and intervention, the existing beliefs and preferences that people have may be leveraged and exploited by AI-assisted systems in promoting misinformation, obtaining unfair profits, and encouraging biased decision-making. Similarly, search systems need to be transparent to users in terms of why and how the search results are generated and make the search results scrutable to users. Systems should also inform users of the potential risks and biases associated with personalized search results, such as confirmation bias, framing effects and echo chamber effects, and offer proactive support to help users avoid or mitigate the negative effects of biases triggered by retrieved results, search recommendations, and users' own previous experiences. The ultimate goal is that users with different backgrounds, existing beliefs, and knowledge base should have equal chance of achieving desired or optimal outcomes, regardless of their individual vulnerability to varying cognitive and perceptual biases in search interactions.

Inspired by equal-odds fairness measures in machine learning (ML) research (cf. Hardt et al., 2016), we write the human-side debiasing or fairness goal as follows:

$$P(Y = Y^* \mid M = 0,\ A = a) = P((Y = Y^* \mid M = 1,\ \ A = a) \tag{7.2}$$

where Y^* refers to the desired or accessible optimal outcome of an individual or group, given the nature of search intentions and motivating task. A represents the set of general contextual attributes that are not directly related to human biases. M indicates if the individual is part of the protected group that is more vulnerable to certain user biases. Note that as it is introduced in previous chapters, different cognitive and perceptual biases may involve different contextual triggers and behavioral impacts. Thus, the associated risks may need to be assessed separately with individual functions. In contrast to algorithmic bias research in AI and ML, users' membership in high risk of human bias category (i.e., protected group) is less likely to be predefined and may need to be inferred from user traits and contextual triggers identified in real-time information seeking and search episodes.

In addition, IR and recommender systems need to be hold accountable when they are used in making automatic critical decisions about different aspects of human-information interaction and everyday life in general, such as approving home loans, generating clinical recommendations based on health records, making hiring decisions, and retrieving a family doctor. The accountability assessment on IR algorithms should include the real-time evaluation of the potential risks of triggering and exploiting factors associated with bounded rationality in making obscure and harmful decisions. This assessment should also cover the *explanations* that search systems provide for justifying recommendations. Systems need to provide explanations that are consistent with how the algorithms *actually* generate real-time recommendations and what features and user information were utilized in the process, rather than simply offering a plausible story that increases the user's acceptance and trust of recommendations by confirming their existing beliefs, biases, and expectations regarding the recommendation mechanism.

Compared to the observable impacts of problematic algorithms on discrimination and bias, the interaction between biased algorithms and boundedly rational users and the associated consequences are usually difficult to characterize, predict, and regulate. Extending the mainstream definition of algorithmic FATE in IR (e.g., Culpepper et al., 2018), we argue that the next-generation AI-assisted search systems should be designed and encouraged to not only monitor and address algorithmic biases (e.g., enhancing fair exposure of documents from different content generators and with different perspectives and political views) but also be transparent about and proactively address the existing problems and potential risks associated with human biases and heuristics in search interaction and judgments of information items. Under the effect of certain cognitive biases and mental shortcuts, users may make local satisficing decisions that may contradict with their goals and tasks behind whole-session interactions. With respect to evaluation, systems need to be assessed in terms of both enhancing algorithmic fairness and transparency and predicting and addressing the potential undesired outcomes caused by human biases and heuristics. Human bias mitigation could be carried out through re-adjusting query recommendations and learning to rank algorithms, or actively reminding users of the possible biases they might have, such as focusing on a narrowed scope of item types or only clicking documents that represent the one single perspective on a controversial topic. Achieving this extended version of FATE in IR will require the integration of insights from data-driven IR experiments, bounded rational research, as well as user interaction design. In practical applications and regulations, the extended FATE approach will go way beyond intelligent search systems and retrieval algorithms themselves and involve a *collective social practice* consisting of actors, forums and platforms, shared beliefs and norms, performativity, as well as regulations and sanctions in broad sociotechnical systems (Johnson, 2021; König, 2020).

7.6 Summary

Chapter 7 brings together the insights from bounded rationality research from multiple disciplines and formal modeling and evaluation in IR and discusses the open problems and new directions under our behavioral economics research agenda. Specifically, based on the research gaps identified in Chap. 6, we take a step forward and discuss more specific problems that need to be addressed in bias-aware IR under three broad sub-areas: Characterizing bounded rationality in IR, developing bias-aware interactive search systems, and bias-aware evaluation. We introduce different ways in which researchers could incorporate the knowledge of bounded rationality into formal user models and different modalities of search systems (especially conversational information seeking and search). Also, we connect our research on bias-aware IR to a broader definition of FATE and present our vision of BITS system, which considers and addresses the negative effects of both algorithmic biases and human biases in human-information interaction and critical decision-making under uncertainty. We hope that the ideas and questions presented in this chapter could encourage future students and researchers to further explore the specific problems and methodological challenges in bias-aware user modeling, system design, and FATE-based system evaluation and include boundedly rational users in the studies of IR and human-AI interaction in general.

References

Agarwal, A., Wang, X., Li, C., Bendersky, M., & Najork, M. (2019). Addressing trust bias for unbiased learning-to-rank. In *The world wide web conference* (pp. 4–14). ACM. https://doi.org/10.1145/3308558.3313697

Ai, Q., Bi, K., Luo, C., Guo, J., & Croft, W. B. (2018). Unbiased learning to rank with unbiased propensity estimation. In *The 41st international ACM SIGIR conference on research & development in information retrieval* (pp. 385–394). ACM. https://doi.org/10.1145/3209978.3209986

Ai, Q., Yang, T., Wang, H., & Mao, J. (2021). Unbiased learning to rank: Online or offline? *ACM Transactions on Information Systems (TOIS), 39*(2), 1–29. https://doi.org/10.1145/3439861

Azzopardi, L. (2021). 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* (pp. 27–37). ACM. https://doi.org/10.1145/3406522.3446023

Büschel, W., Mitschick, A., & Dachselt, R. (2018). Here and now: Reality-based information retrieval: Perspective paper. In *Proceedings of the 2018 ACM SIGIR conference on human information interaction & retrieval* (pp. 171–180). https://doi.org/10.1145/3176349.3176384

Chen, R. C., Ai, Q., Jayasinghe, G., & Croft, W. B. (2019). Correcting for recency bias in job recommendation. In *Proceedings of the 28th ACM international conference on information and knowledge management* (pp. 2185–2188). ACM. https://doi.org/10.1145/3357384.3358131

Chen, N., Zhang, F., & Sakai, T. (2022). Constructing better evaluation metrics by incorporating the anchoring effect into the user model. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval* (pp. 2709–2714). ACM. https://doi.org/10.1145/3477495.3531953

Craswell, N., Zoeter, O., Taylor, M., & Ramsey, B. (2008). An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 87–94). ACM. https://doi.org/10.1145/1341531.1341545

Culpepper, J. S., Diaz, F., & Smucker, M. D. (2018). Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in Lorne (SWIRL 2018). In *ACM SIGIR forum* (Vol. 52, pp. 34–90). ACM. https://doi.org/10.1145/3274784.3274788

Deldjoo, Y., Trippas, J. R., & Zamani, H. (2021). Towards multi-modal conversational information seeking. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval* (pp. 1577–1587). ACM. https://doi.org/10.1145/3404835.3462806

Dungs, S., & Fuhr, N. (2017). Advanced hidden Markov models for recognizing search phases. In *Proceedings of the ACM SIGIR international conference on theory of information retrieval* (pp. 257–260). ACM. https://doi.org/10.1145/3121050.3121090

Eickhoff, C. (2018). Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining* (pp. 162–170). ACM. https://doi.org/10.1145/3159652.3159654

Ge, Y., Zhao, S., Zhou, H., Pei, C., Sun, F., Ou, W., & Zhang, Y. (2020). Understanding echo chambers in e-commerce recommender systems. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (pp. 2261–2270). https://doi.org/10.1145/3397271.3401431

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science, 349*(6245), 273–278. https://doi.org/10.1126/science.aac6076

Gilboa, I., & Schmeidler, D. (1995). Case-based decision theory. *The Quarterly Journal of Economics, 110*(3), 605–639. https://doi.org/10.2307/2946694

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems, 29*.

Hendahewa, C., & Shah, C. (2013). Segmental analysis and evaluation of user focused search process. In *Proceedings of the 2013 12th international conference on machine learning and applications* (Vol. 1, pp. 291–294). IEEE. https://doi.org/10.1109/ICMLA.2013.59

Jansen, M., Bos, W., Van Der Vet, P., Huibers, T., & Hiemstra, D. (2010). TeddIR: Tangible information retrieval for children. In *Proceedings of the 9th international conference on interaction design and children* (pp. 282–285). https://doi.org/10.1145/1810543.1810592

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 133–142). ACM. https://doi.org/10.1145/775047.775067

Joachims, T., Swaminathan, A., & Schnabel, T. (2017). Unbiased learning-to-rank with biased feedback. In *Proceedings of the tenth ACM international conference on web search and data mining* (pp. 781–789). ACM. https://doi.org/10.1145/3018661.3018699

Johnson, D. G. (2021). Algorithmic accountability in the making. *Social Philosophy and Policy, 38*(2), 111–127. https://doi.org/10.1017/S0265052522000073

Jones, E. E., Rock, L., Shaver, K. G., Goethals, G. R., & Ward, L. M. (1968). Pattern of performance and ability attribution: An unexpected primacy effect. *Journal of Personality and Social Psychology, 10*(4), 317–340. https://doi.org/10.1037/h0026818

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review, 93*(5), 1449–1475. https://doi.org/10.1257/000282803322655392

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

König, P. D. (2020). Dissecting the algorithmic leviathan: On the socio-political anatomy of algorithmic governance. *Philosophy & Technology, 33*(3), 467–485. https://doi.org/10.1007/s13347-019-00363-w

Lagun, D., Hsieh, C. H., Webster, D., & Navalpakkam, V. (2014). Towards better measurement of attention and satisfaction in mobile search. In *Proceedings of the 37th international ACM SIGIR*

conference on research & development in information retrieval (pp. 113–122). ACM. https://doi.org/10.1145/2600428.2609631

Lankton, N. K., & McKnight, H. D. (2012). Examining two expectation disconfirmation theory models: Assimilation and asymmetry effects. *Journal of the Association for Information Systems, 13*(2), 88–115. https://doi.org/10.17705/1jais.00285

Law, E., Yin, M., Goh, J., Chen, K., Terry, M. A., & Gajos, K. Z. (2016). Curiosity killed the cat, but makes crowdwork better. In *Proceedings of the 2016 ACM SIGCHI conference on human factors in computing systems* (pp. 4098–4110). ACM. https://doi.org/10.1145/2858036.2858144

Leon, K., Walker, W., Lim, Y., Penman, S., Colombo, S., & Casalegno, F. (2019). Tangible map: Designing and assessing spatial information retrieval through a tactile interface. In *International conference on human-computer interaction* (pp. 329–340). Springer. https://doi.org/10.1007/978-3-030-22636-7_24

Li, H. (2011). A short introduction to learning to rank. *IEICE Transactions on Information and Systems, 94*(10), 1854–1862. https://doi.org/10.1587/transinf.E94.D.1854

Li, Y., & Belkin, N. J. (2008). A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management, 44*(6), 1822–1837. https://doi.org/10.1016/j.ipm.2008.07.005

Lin, J., Liu, W., Dai, X., Zhang, W., Li, S., Tang, R., He, X., Hao, J., & Yu, Y. (2021). A graph-enhanced click model for web search. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval* (pp. 1259–1268). ACM. https://doi.org/10.1145/3404835.3462895

Liu, J. (2021). Deconstructing search tasks in interactive information retrieval: A systematic review of task dimensions and predictors. *Information Processing & Management, 58*(3), 102522.

Liu, J. (2022). Toward Cranfield-inspired reusability assessment in interactive information retrieval evaluation. *Information Processing & Management, 59*(5), 103007. https://doi.org/10.1016/j.ipm.2022.103007

Liu, J., & Han, F. (2020). 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* (pp. 1141–1150). ACM. https://doi.org/10.1145/3397271.3401085

Liu, J., & Shah, C. (2019). Investigating the impacts of expectation disconfirmation on web search. In *Proceedings of the 2019 ACM SIGIR conference on human information interaction and retrieval* (pp. 319–323). ACM. https://doi.org/10.1145/3295750.3298959

Liu, J., & Shah, C. (2022). Leveraging user interaction signals and task state information in adaptively optimizing usefulness-oriented search sessions. In *Proceedings of the 22nd ACM/IEEE joint conference on digital libraries* (pp. 1–11). ACM. https://doi.org/10.1145/3529372.3530926

Liu, J., & Yu, R. (2021). State-aware meta-evaluation of evaluation metrics in interactive information retrieval. In *Proceedings of the 30th ACM international conference on information & knowledge management* (pp. 3258–3262). ACM. https://doi.org/10.1145/3459637.3482190

Liu, M., Mao, J., Liu, Y., Zhang, M., & Ma, S. (2019). Investigating cognitive effects in session-level search user satisfaction. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 923–931). ACM. https://doi.org/10.1145/3292500.3330981

Liu, J., Sarkar, S., & Shah, C. (2020). Identifying and predicting the states of complex search tasks. In *Proceedings of the 2020 ACM SIGIR conference on human information interaction and retrieval* (pp. 193–202). ACM. https://doi.org/10.1145/3343413.3377976

Luo, J., Zhang, S., & Yang, H. (2014). Win-win search: Dual-agent stochastic game in session search. In *Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval* (pp. 587–596). ACM. https://doi.org/10.1145/2600428.2609629

Maddalena, E., Basaldella, M., De Nart, D., Degl'Innocenti, D., Mizzaro, S., & Demartini, G. (2016). Crowdsourcing relevance assessments: The unexpected benefits of limiting the time to judge. In *Proceedings of the AAAI conference on human computation and crowdsourcing* (Vol. 4, pp. 129–138) https://ojs.aaai.org/index.php/HCOMP/article/view/13284

Ma-Kellams, C. (2020). Cultural variation and similarities in cognitive thinking styles versus judgment biases: A review of environmental factors and evolutionary forces. *Review of General Psychology, 24*(3), 238–253. https://doi.org/10.1177/1089268019901270

Mao, J., Luo, C., Zhang, M., & Ma, S. (2018). Constructing click models for mobile search. In *The 41st international ACM SIGIR conference on research & development in information retrieval* (pp. 775–784). ACM. https://doi.org/10.1145/3209978.3210060

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Ong, K., Järvelin, K., Sanderson, M., & Scholer, F. (2017). Using information scent to understand mobile and desktop web search behavior. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval* (pp. 295–304). ACM. https://doi.org/10.1145/3077136.3080817

Posch, L., Bleier, A., Lechner, C. M., Danner, D., Flöck, F., & Strohmaier, M. (2019). Measuring motivations of crowdworkers: The multidimensional crowdworker motivation scale. *ACM Transactions on Social Computing, 2*(2), 1–34. https://doi.org/10.1145/3335081

Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J., & Vukovic, M. (2011). An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *Proceedings of the international AAAI conference on web and social media* (Vol. 5, pp. 321–328) https://ojs.aaai.org/index.php/ICWSM/article/view/14105

Roitero, K., Checco, A., Mizzaro, S., & Demartini, G. (2022). Preferences on a budget: Prioritizing document pairs when crowdsourcing relevance judgments. In *Proceedings of the ACM web conference* (pp. 319–327). ACM. https://doi.org/10.1145/3485447.3511960

Rosenthal, R. (1976). *Experimenter effects in behavioral research*. Irvington Publishers.

Scholer, F., Kelly, D., Wu, W. C., Lee, H. S., & Webber, W. (2013). The effect of threshold priming and need for cognition on relevance calibration and assessment. In *Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval* (pp. 623–632). ACM. https://doi.org/10.1145/2484028.2484090

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). *Towards a standard for identifying and managing bias in artificial intelligence*. NIST special publication, 1270.

Sekulić, I., Aliannejadi, M., & Crestani, F. (2021). Towards facet-driven generation of clarifying questions for conversational search. In *Proceedings of the 2021 ACM SIGIR international conference on theory of information retrieval* (pp. 167–175). ACM. https://doi.org/10.1145/3471158.3472257

Shen, S., Hu, B., Chen, W., & Yang, Q. (2012). Personalized click model through collaborative filtering. In *Proceedings of the fifth ACM international conference on web search and data mining* (pp. 323–332). ACM. https://doi.org/10.1145/2124295.2124336

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics, 69*(1), 99–118. https://doi.org/10.2307/1884852

Speicher, M., Both, A., & Gaedke, M. (2015). SOS: Does your search engine results page (SERP) need help? In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 1005–1014). ACM. https://doi.org/10.1145/2702123.2702568

Thaler, R. H. (2016). Behavioral economics: Past, present, and future. *American Economic Review, 106*(7), 1577–1600. https://doi.org/10.1257/aer.106.7.1577

Thomas, P., Czerwinksi, M., McDuff, D., & Craswell, N. (2021). Theories of conversation for conversational IR. *ACM Transactions on Information Systems (TOIS), 39*(4), 1–23. https://doi.org/10.1145/3439869

Thomas, P., Kazai, G., White, R., & Craswell, N. (2022). The crowd is made of people: Observations from large-scale crowd labelling. In *Proceedings of the international ACM SIGIR*

conference on human information interaction and retrieval (pp. 25–35). ACM. https://doi.org/10.1145/3498366.3505815

Trippas, J. R., Spina, D., Cavedon, L., Joho, H., & Sanderson, M. (2018). Informing the design of spoken conversational search: Perspective paper. In *Proceedings of the 2018 ACM SIGIR conference on human information interaction & retrieval* (pp. 32–41). ACM. https://doi.org/10.1145/3176349.3176387

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics, 106*(4), 1039–1061. https://doi.org/10.2307/2937956

Vardasbi, A., de Rijke, M., & Markov, I. (2021). Mixture-based correction for position and trust bias in counterfactual learning to rank. In *Proceedings of the 30th ACM international conference on information & knowledge management* (pp. 1869–1878). ACM. https://doi.org/10.1145/3459637.3482275

Venkatesh, V., & Goyal, S. (2010). Expectation disconfirmation and technology adoption: Polynomial modeling and response surface analysis. *MIS Quarterly, 34*(2), 281–303. https://doi.org/10.2307/20721428

Weber, R. A., & Camerer, C. F. (2006). "Behavioral experiments" in economics. *Experimental Economics, 9*(3), 187–192. https://doi.org/10.1007/s10683-006-9121-5

White, R. (2013). Beliefs and biases in web search. In *Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval* (pp. 3–12). ACM. https://doi.org/10.1145/2484028.2484053

Yamamoto, Y., & Yamamoto, T. (2018). Query priming for promoting critical thinking in web search. In *Proceedings of the 2018 international ACM SIGIR conference on human information interaction & retrieval* (pp. 12–21). ACM. https://doi.org/10.1145/3176349.3176377

Yan, R., Li, J., & Yu, Z. (2022). Deep learning for dialogue systems: Chit-chat and beyond. *Foundations and Trends in Information Retrieval, 15*(5), 417–589. https://doi.org/10.1561/1500000083

Zhai, C. (2016). Towards a game-theoretic framework for text data retrieval. *IEEE Data Engineering Bulletin, 39*(3), 51–62.

Zhang, Y., & Zhai, C. (2016). A sequential decision formulation of the interface card model for interactive IR. In *Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval* (pp. 85–94). ACM. https://doi.org/10.1145/2911451.2911543

Zhang, Y., Chen, X., Ai, Q., Yang, L., & Croft, W. B. (2018). Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th ACM international conference on information and knowledge management* (pp. 177–186). ACM. https://doi.org/10.1145/3269206.3271776