Boundedly Rational Searchers Interacting with Medical Misinformation: Characterizing Context-Dependent Decoy Effects on Credibility and Usefulness Evaluation in Sessions

Jiqun Liu*
School of Library and Information Studies
The University of Oklahoma
Norman, Oklahoma, USA
jiqunliu@ou.edu

Jiangen He
The University of Tennessee - Knoxville
Knoxville, Tennessee, USA
jiangen@utk.edu

Abstract

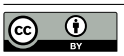
Characterizing users' judgments and interactions with search engine result pages (SERPs) has been a central theme in Interactive Information Retrieval (IIR) evaluation. In contrast to the perfect rationality assumptions underpinning most existing formal models, people are *boundedly rational* and are subject to the influence of systematic cognitive biases. To enhance the psychological foundation of user models and better understand the *in-situ* decisions of boundedly rational users, our between-subject crowdsourcing experiment explored *Decoy Effect* on users' vulnerability to COVID treatment misinformation, which causes enduring impacts on people's personal health management and wellbeing, and examined the extent to which this effect is moderated by contextual factors and user characteristics. Our results, derived from 540 participants and 2,160 valid SERP evaluation records, indicate that: 1) users' interactions with decoy results may increase their vulnerability to medical misinformation in usefulness and credibility judgments; 2) the size of decoy effect is conditioned by users' prior knowledge and the rank position of decoy results. This research empirically reveals the impact of decoy results on users' *context-dependent* preferences on ranked search results under varying topics and conditions of medical information evaluation. More broadly, it demonstrates the value of representing and modeling users interacting with information as boundedly rational agents and serves as a step forward towards achieving the goal of truly human-centered IIR.

CCS Concepts

• **Information systems → Users and interactive retrieval**.

ACM Reference Format:
Jiqun Liu and Jiangen He. 2025. Boundedly Rational Searchers Interacting with Medical Misinformation: Characterizing Context-Dependent Decoy Effects on Credibility and Usefulness Evaluation in Sessions. In *2025 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '25), March 24–28, 2025, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3698204.3716453

*Corresponding author.



CHIIR '25, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-1290-6/25/03
https://doi.org/10.1145/3698204.3716453

1 Introduction

Characterizing users' preferences and *in-situ* judgments has been a central theme in Interactive Information Retrieval (IIR) evaluation. In contrast to the (over)simplified assumptions underpinning a variety of formal user models and evaluation metrics, people are *boundedly rational* and often influenced by a series of systematic *cognitive biases* [16, 33, 73]. According to Kahneman's *Two-Systems* framework [cf. 32], decision makers' *in-situ* preferences over multiple options are often established unconsciously and quickly, and are usually influenced by their cognitive biases, accessible memories of experiences and heuristics. The operations of *System 1* behind judgments diminishes individuals' motivations for pursuing careful comparison of options and accurate estimation of gains and costs. Consequently, model predictions built upon *perfect rationality* assumptions often deviate significantly from users' actual decisions and retrospective evaluations [43, 46] as individuals' preferences tend to be *contextual-dependent*. While mental shortcuts may accelerate information processing under a variety of everyday-life tasks, they could undermine our ability to accurately evaluate the *usefulness* and *credibility* of information and also influence decisions without our awareness [12, 43]. The negative effect of cognitive limits, biased results, and misinformation become even more severe and ubiquitous as immediate answers and system-generated human-like responses often appear to be readily available through diverse online sources, especially under the wide and fast-growing applications of large language models (LLMs) [7, 21, 22, 74].

To better understand users' judgments and predict their behaviors more accurately (especially in high-stakes tasks, such as medical information search), it is essential to investigate the impacts of cognitive biases and their interactions with potential moderating factors, such as users' prior behavior, topical knowledge, and document rank positions [44, 62, 79, 87]. However, due to the difficulty of empirically characterizing cognitive biases and the potential computational complexity of modeling them, the role of *cognitive biases* is often abstracted out of formal models in IR evaluation and thus remains understudied (with a few exceptions [e.g. 24, 25, 46, 49]), especially when compared to rapidly growing research on *algorithmic biases* and system-oriented fairness [10, 37]. This gap in research hinders the further development of user models and human-centered systems and motivates our research reported here.

Specifically, our study explores the role of human biases in *interactive session search* consisting of multiple query-SERP iterations and focuses on *Decoy Effect* [80, 81], a cognitive bias that has been

empirically confirmed by behavioral economics and cognitive psychology experiments involving diverse domains, but has not been thoroughly examined in information access and human-centered computing research [42]. Decoy effect, which contradicts the predictions of rational models and *expected utility* theory [53], refers to the phenomenon in which the presence of an inferior option can influence an individual's preference between two similar options in decision-making [73]. This variation in *context-dependent* preference could not be characterized with simplified gain-cost analysis or traditional utility framework as there is no change in the options or associated mathematical expectations [46].

Studying decoy effects has important practical values confirmed by both experimental research and applications. In marketing, comprehending how decoy options steer preferences empowers businesses to design more effective pricing strategies and product presentations, thereby optimizing consumer choices and enhancing profitability. Moreover, in the realm of public policy, recognizing the impact of decoy effects enables the design of interventions that subtly nudge individuals toward decisions that align with societal objectives, fostering healthier lifestyles or sustainable practices. In essence, studying decoy effects serves as a gateway to deciphering the nuances of human choice, influencing fields ranging from economics to neuroscience, and guiding ethical considerations in their utilization. In IR, the knowledge of decoy effect could allow researchers to better understand the preferences and judgments of real-life users (as opposed to simulated agents) on information objects and predict their behaviors more accurately [43].

To empirically demonstrate the impact of decoy items and examine the conditions that trigger the effect, we conducted a crowdsourcing *between-subject* experiment where we investigated participants' judgments on COVID-related medical information credibility and usefulness under pre-designed *decoy* and *non-decoy* SERP scenarios. In particular, we studied how decoy search results affect a user's *in-situ vulnerability* to COVID misinformation, which in this study measures how likely the user would click, read, and assign a high credibility and/or usefulness score to an misinformation page about COVID treatment under a particular condition of search interaction (e.g. decoy result present or absent). Furthermore, to explore the moderating effect of contextual factors, we examined the extent to which the impact of decoy results varies according to different motivating questions, rank positions, users' levels of topical knowledge, and *depth* of interactions with decoy options (e.g. viewing on SERP only, clicking, judging and rating).

Our research takes a deep dive into the decoy effect and demonstrates that: 1) overall, users' *in-situ* credibility judgments are more vulnerable to negative decoy effects compared to usefulness judgments; 2) deeper interactions with decoy options (e.g. clicking on decoy results and giving a low rating on the results) are associated with more significant impacts on medical information judgments (e.g. higher usefulness and credibility ratings on medical misinformation); 3) the effect size of decoy varies across different levels of user domain knowledge and rank positions on SERPs. The main contributions of our work are threefold:

- We comprehensively investigate the effect of decoy items in medical information judgments within *multi-query/interactive session* contexts and present direct empirical evidence on the impact of decoy results on users' search behavior and vulnerability to COVID misinformation under varying rank positions, levels of topical knowledge, and SERP conditions. Our findings can inform the design and assessment of cognitive debiasing techniques, especially in the context of medical information evaluation and personal health decision-making.
- Our study bridges the insights on cognitive biases from behavioral economics and the research on user judgments in IR, empirically confirms the effect of decoy options in a new context, and methodologically presents an effective experimental setup that could be reused and replicated in examining cognitive biases in future studies.
- Overall, this study demonstrates the value of leveraging the knowledge about human cognitive biases in studying and reframing IR problems and enhances our understanding of real-life searchers' context-dependent preferences and interactions with online medical information. This approach enriches user modeling and human-centered system evaluation in medical information access by adding new bias dimensions to current frameworks.

2 Related Work

This section will present the fundamental concepts and interdisciplinary progresses underpinning our work. It starts with incorporating *boundedly rationality perspective* into the problem of user modeling, and then emphasize the impact of *decoy effect in decision-making*, which is the focus of our experiment. Lastly, this section discusses previous advances and open challenges in characterizing and predicting *user judgments* in human-centered IR evaluation.

2.1 Understanding Boundedly Rational Users

Users as people interacting with information are not perfectly rational (e.g. always pursuing optimized utility; having full access to information and unlimited cognitive resources for computing gains and costs) as it is implicitly assumed in most user models. Instead, they are *boundedly rational* and are subject to the influence of cognitive biases, emotions, mental shortcuts and heuristics [1, 42, 43, 52], which usually lead to systematic deviations of users' actual preferences and behaviors from the predictions of simplified formal models. Apart from the decades of behavioral experiments from other disciplines [e.g. 73], some recent studies from information seeking and search communities have also attempted to describe the impact of cognitive biases on search tactics, information evaluation and information use in decision-making [5, 25, 46, 78]. However, many of the biases discussed in previous IR research were examined in *topical relevance* judgment and *ad hoc* retrieval contexts only. Therefore, it is still unclear how these biases are triggered in sessions of information search and evaluation, and how the impact of cognitive biases vary across different aspects of search interactions (e.g. clicking, browsing and reading, *in-situ* judgments and preferences) [6]. Consequently, the disconnection between the knowledge learned from describing biases and the computational models of user judgments developed in offline experiments still remains [42]. Addressing this problem will allow researchers to identify system and behavioral features that could be used for predicting the occurrence of cognitive biases, and also facilitate the design of proactive

nudging and interventions for mitigating the negative effect of biases.

2.2 Decoy Effect in Decision-Making

Decoy effect, also known as *asymmetric dominance* effect, has been extensively studied and recognized in the fields of behavioral economics and decision making [59, 80]. [75] first introduced decoy effect through experiments and demonstrated that the presence of a decoy option can significantly alter individuals' preferences and the perceived attractiveness of the options being considered, even though there is no change to these options. For instance, a decision between "**A**: no journal subscription with no fee" and "**B**: subscription with $20 annual fee" maybe difficult to predict. However, when adding a third choice, "**C**: subscription with $30 annual fee", as a decoy option, the option **B** may appear to be more attractive, even though the decoy option itself is not chosen.

Following [75], subsequent studies have investigated decoy effect and confirmed its impacts in a broad range of domains and decision-making scenarios, such as consumer purchasing decisions [30, 81, 85], medical treatment choices [11, 71], and children's behavior [91]. Going beyond explicit behavioral changes, [29] explored the underlying neural correlates of decoy effect and found that choice sets with decoys activated the occipital gyrus and deactivated the inferior parietal gyrus. One potential explanation for the decoy effect is built upon the concept of *salience*, or the extent to which an option stands out from others. When a decoy option as a *reference point* is presented, the relative salience of the original options may be altered, causing the increase in *perceived attractiveness* of one of the options [20, 68]. Evidences of decoy effect directly contradicts the "context-invariant" assumption in Economics and offers an alternative explanation on seemingly irrational choices.

Differing from purchasing and simplified decision-making tasks in behavioral experiments, health and medical information evaluation in Web search involves a set of mini-decision moments (e.g. viewing, clicking, judgment) and is closely associated with a variety of contextual factors, such as the rank position of search results, system layouts, topics, and users' topical knowledge [17, 51, 66, 90]. Examining decoy effect and its correlations with search contexts will enhance our understanding of user judgments and search decisions. Meanwhile, it also requires a carefully designed experimental setting for sorting out the mixed effects from interrelated variables.

2.3 User Judgment in Information Retrieval Evaluation

Understanding user judgments and preferences has been a main component of IR evaluation research [26], especially under the impact of widespread online misinformation [70, 86]. Among different types of judgments and labels, *relevance judgment*, which measures the extent to which a document is topically relevant to a query issued, serves as the basis for decades of system-centered IR evaluation experiments [26, 77]. Offline evaluation measures built upon *query-document relevance* determines the criteria for differentiating good-performing retrieval systems from the other ones and have been employed for training and evaluating ranking algorithms [36]. While relevance-based Cranfield paradigm facilitates the rapid development of *ad hoc* factual retrieval techniques [77], it is insufficient for evaluating and enhancing multi-query information seeking episodes triggered by complex search tasks [9, 58]. Overcoming this limitation requires further research on at least two aspects: 1) going beyond topical relevance and characterizing other dimensions of user judgments; 2) investigating behavior and judgments in *sessions* [2, 40], rather than *one-query-one-response* contexts [13].

Regarding the first aspect, some IIR researchers have investigated several user-centered dimensions of IR evaluation, such as *usefulness* judgment [19, 41], *credibility* and *trustworthiness* judgment [28, 34, 39, 82, 84], and retrospective satisfaction and engagement rating [2, 56, 88], especially under the potential threats of online misinformation and disinformation [e.g. 57, 61, 76, 89, 92]. Furthermore, [50] leveraged usefulness feedback in enhancing search recommendation algorithms and helping users complete search tasks sooner, offering a new perspective for IR evaluation.

Regarding the second aspect, while progresses have been made on understanding multiple dimensions of user judgment, most of the experiments have been conducted in either *ad hoc* retrieval settings [e.g. 18] or evaluation-only contexts (isolated document judgment phases without enabling other actions of search processes) [e.g. 35, 67]. [25] has examined cognitive biases (including decoy effect) in Crowdsourcing tasks and measured their effects on user judgments of documents. However, search activities and session contexts were abstracted out from the controlled evaluation process, leading to difficulties in characterizing the interaction between decoy options and other factors. To address this limitation, one needs to overcome a methodological challenge: Developing a reasonably controlled environment for triggering and observing the behavioral effect of cognitive biases (which users themselves may not be aware of) and also maintaining certain levels of authenticity of the simulated search sessions and user experiences.

3 Research Questions

To address the gaps above, we conducted a *between-subject* Crowdsourcing experiment which simulated a search session experience consisting of four query-SERP combinations for each sub-topic. Using a between-subjects design helps to minimize potential carryover effects and learning biases that can occur when participants are exposed to multiple conditions. Our work seeks to answer following two research questions (RQs):

- **RQ1:** To what extent do users' credibility and usefulness judgments vary across different levels of users' interactions with decoy search results in medical information evaluation?
- **RQ2:** How do search contextual factors moderate decoy effects on users' judgments and search behavior?

Instead of simply comparing between-group differences in users' ratings, under RQ1, we took a step forward and examined what level of interaction with decoy results (e.g. viewing the snippet on SERP, viewing the result, giving the result a low rating) can actually *trigger* decoy effect on user judgments. In RQ2, we explored how the impact of user characteristics and SERP factors (e.g. rank position) varies when a decoy search result is presented.

We adopted a Crowdsourcing experiment design instead of a traditional controlled laboratory experiment mainly for following reasons: 1) Crowdsourcing experiment has been widely applied and confirmed to be a useful method for IR evaluation and allow researchers to recruit a large group of participants from diverse backgrounds (especially outside college students population) [38, 63, 64]; 2) Crowdsourcing platforms can enable a fast recruitment process and reduce the possibility of delays in recruitment and associated analyses, which in turn can help control the budget for user-centered IR evaluation [48]; 3) Our evaluation experiment (including search interaction, judgment, and response submission) has a straightforward and well-defined procedure and has also been tested in pilot feasibility studies, and thus does not rely on a resource-consuming laboratory setting. Our experiment was approved by the University of Oklahoma's Institutional Review Board (**IRB #**: 13527).

4 Methodology

This section introduces the study design and analysis methods we employed for answering the RQs presented above.

4.1 Participants

Our participants/crowdworkers were recruited through Prolific [1], a crowdsourcing platform for online behavioral research. Five Prolific's prescreeners have been applied to this study: 1) Native English speakers, 2) Without professional education in medicine or related areas, 3) at least 50% of studies for which the participant have been approved, and 4) No literacy difficulty. 40 participants completed the pilot study and 617 completed the full experiment: 157 were in Control A, 148 were in Control B, 158 were in Treatment Group A, and 154 were in Treatment Group B. 325 (52.7%) participants were female; 284 (46.0%) were male; 8 (1.3%) preferred not to say about their gender. Their ages ranged from 18 to 83 years (M=27.8, SD=8.5). 253 (41.0%) participants had an undergraduate degree, 179 (29.0%) had a high school diploma, 118 (17.8%) had a graduate degree, and 60 (9.7%) had a technical/community college degree. 374 (60.6%) participants were from the United Kingdom, 156 (25.3%) were from the United States, 52 (8.4%) were from Canada, and 29 (4.7%) were from Australia. The majority of participants were students (453, 73.4%). 395 participants (64.0%) were employed (249 full-time employees and 146 part-time employees), 100 (16.2%) were unemployed or not in paid work, and 118 (19.1%) were in other employment statuses. Participants were informed of the task and experimental procedure prior to signing up for the study.

4.2 Selection of Topics

To properly simulate the experience of evaluating pages in *sessions*, we selected three COVID treatment related topics/motivating questions that require certain level of expertise and acquisiton of new information and knowledge for most users, including Monoclonal Antibodies, ACE and ARB, and Hydroxychloroquine in Table 1. We narrowed down the scope to COVID treatment because 1) it is a critical medical information topic that could cause significant, long-lasting social and economic impact, and 2) the narrowed topic scope helps reduce the potential behavioral impact of domain variations and thereby control the hidden contextual effect on decoy impact measurements.

Table 1: Topics for user studies.

Topics	Questions
1. Monoclonal antibodies	Can monoclonal antibodies cure COVID-19?
2. ACE and ARBs	Can ACE and ARBs worsen COVID-19?
3. Hydroxychloroquine	Can hydroxychloroquine treat COVID-19?

Regarding the three questions/treatment options, monoclonal antibodies are laboratory-made molecules designed to mimic the immune system's ability to fight off harmful pathogens, such as viruses. For COVID-19 treatment, monoclonal antibodies have been developed to target and neutralize the SARS-CoV-2 virus, which causes COVID-19. These antibodies were proposed and administered as a treatment option for individuals who have tested positive for the virus and are at high risk of developing severe disease. ACE inhibitors and ARBs are classes of medications commonly used to manage conditions related to high blood pressure (hypertension) and heart failure. In the context of COVID-19, there was growing interest in these medications due to their interaction with the renin-angiotensin-aldosterone system (RAAS), which is also involved in regulating blood pressure and fluid balance. Hydroxychloroquine is an antimalarial drug that was explored as a potential treatment for COVID-19 early in the pandemic. It gained attention due to its in vitro antiviral properties against SARS-CoV-2. However, subsequent clinical trials and studies found inconsistent and inconclusive results regarding its effectiveness in treating COVID-19.

We selected these particular COVID subtopics from the TREC Health Misinformation track dataset [2], for two main **reasons**: 1) the tasks/questions under these topics have YES/NO answers annotated by external expert assessors, which can be used as relible ground truth for judging information credibility and differentiating correct information from misinformation, 2) compared to general COVID health information (e.g. COVID vaccination), these subtopics involve specific domain knowledge related to COVID treatment and thus tend to be almost equally unfamiliar to the participants. The answers as ground truth labels helped us differentiate misinformation from correct information. We collected 40 Web pages for each task/question (four predefined queries/SERPs, with ten organic results on each SERP).

4.3 Decoy Design

To study how decoy effect could be triggered by signals on SERPs, we developed decoy search results by manipulating the *title* and *short abstract* of a search result snippet on each SERP under treatment groups. Specifically, to construct the experimental condition under which decoy effect can be measured, under the same COVID treatment topic, the decoy result and target misinformation result on a SERP share the same or very similar subtopics relevant to the task question, which are slightly different from that of the correct information [3]. To frame the decoy result as an inferior option and

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

[2] https://trec-health-misinfo.github.io/2020.html

[3] Although the correct information result does not share exactly the same subtopic with the decoy result, it is still designed to be highly relevant to the task.

enhance the perceived saliency, we enhanced the SERP snippet of *target misinformation results* with more specific *subject or entity* information (e.g. the lab or university that participated in the testing of certain COVID treatments) and/or *statistics* (e.g. total number of participants and researchers involved in the medication experiments; number of experiments conducted under similar conditions). In contrast, the decoy result, without any specific information about the subject or clinical statistics, tend to be perceived as *obviously false*. Based on previous findings on users' common strategies of health and medical information judgments [e.g. 23, 90], our *empirically supported* assumption is that titles and search result snippets with more specific information (e.g. name of the lab, number of participants) and statistics of clinical trials tend to be perceived as high-quality and more trustworthy in medical information evaluation, leading to salient, perceivable difference between target results and decoy options.

To illustrate our decoy design approach, Figure 2 presents example search results from treatment group and control group respectively under the topic of "Hydroxychloroquine". In the treatment group, the decoy and target results share the same subtopic: the treatment effect of hydroxychloroquine on COVID-19 symptoms. Although the correct result shares the same general topic, the specific subtopic or question it focuses on (i.e. hydroxychloroquine and hospitalized patients) is slightly different from that of the other two results. Compared to the decoy result, the title and search result snippet of the target result offers more details on the subject information (e.g. international clinical trial) and statistics information (e.g. primary outcomes assessed at 90 days after randomization). With the decoy result as a low-quality reference point, the target result may be perceived as more credible and reliable. Similarly, in Figure 3, the target result offers more detail about the subject (i.e. researchers from Mount Sinai medical system), whereas the decoy result only presents the conclusion, leading to a relatively higher level of perceived credibility on the target result (misinformation). Also, this decoy effect may cause the deviation of users' attention from the correct information result. We adopted this decoy design approach, aiming to **balance two goals**: 1) creating a reasonable experimental condition to examine the role of decoy effect; 2) maintaining a certain level of authenticity in simulating natural search experience, without introducing unrealistically strong or oversimplified treatments that are only valid in experimental settings.

To control the experimental setting and limit potential mixed effects, no change was introduced to the original main text of Web pages. With all other results controlled, the presence of decoy search result was the *only* difference between a treatment group and the corresponding control group under the same subtopic. In this session-based setting, each treatment group session has four decoy results in total (one on each predefined SERP/query segment).

Ranking and rank-based offline evaluation are a central topic in IR. To examine the impact of *rank position* bias, we developed two decoy effect group: group **A** presents the decoy result before the target misinformation result so that individuals are more likely to encounter and click decoy option first, while group **B** ranks the misinformation result higher than the decoy result. Figure 1 illustrates the design of four experimental conditions. Note that all three-result combinations are ranked at the top three positions on corresponding SERPs. All other seven organic search results,

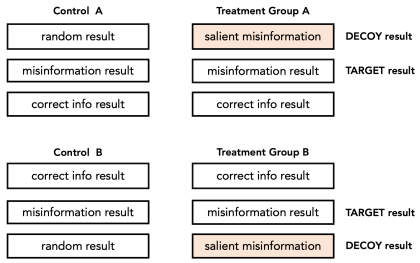


Figure 1: Design of Experimental Conditions. A decoy result was added to EACH SERP in treatment sessions.

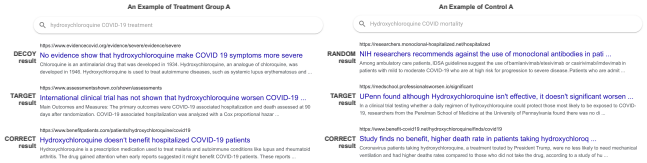


Figure 2: Examples of the first three search results in the Control and Treatment Group. The topic of the examples is "Hydroxychloroquine".

which are less topically relevant compared to the top three results, remained the same under the same query across different experimental conditions.

4.4 Crowdsourcing Experiment Flow

Figure 3 shows the experiment flow of our crowdsourcing experiment, which includes five steps:

(1) Complete an eligibility survey and read the instructions.
(2) Answer questions that measure their prior knowledge on the topic.
(3) Start a search with a query page and click the search button to proceed. A participant searches with four queries.
(4) Browse the result page of each query.
(5) Read the full text of the results. For each result page, a participant has to read the full text of at least three results.
(6) Evaluate the credibility and usefulness of the article. A dialog would pop up after a participant closes the full text of the article and the participant has to give ratings to the article about its credibility and usefulness.
(7) Answer the task question (e.g. "Can monoclonal antibodies cure COVID-19"?).

4.5 Data Collection and Analysis

Our data collection was completed through three sessions on Prolific. First, we conducted a *pilot study* with 40 participants on the topic of Monoclonal Antibodies to test our system, interface, and the clarity of the study procedure. Several revisions to the experiment design were made based on the preliminary results, user-reported confusions, unexpected technical errors, and limitations in study procedures found during the pilot study. Participants were required to complete the four queries for a topic and open at least three articles in the results of each query to proceed to answer the question

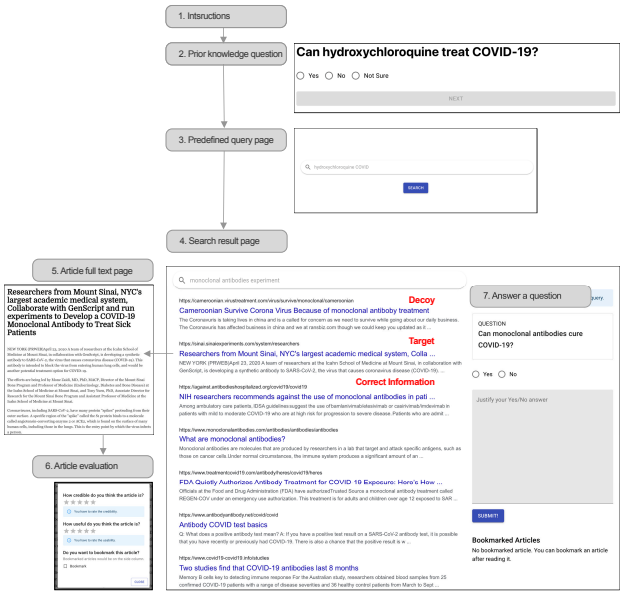


Figure 3: Crowdsourcing Experiment Flow.

and complete their study. To obtain sufficient data on participants' evaluation of credibility and usability, participants are required to rate the credibility and usability of an article after they read the article for the first. Participants can choose to change their ratings after the first time of reading, but it is not required. Therefore, each participant would provide ratings for at least 12 articles.

We collected data from 540 participants in the first data collection phase. 45 participants were assigned to each of 12 conditions (3 topics × 4 treatments). Data from 474 participants were effective and retained after data from 60 participants who spent less than five minutes in the study and 6 participants who provided the same ratings for all the articles. The second data collection session was conducted to collect data from 45 participants. This session was iterative. For each round, the number of participants recruited for each condition was the same before the number of participants reaches 45. The data collection ended when we successfully collected effective data from 45 participants for each condition and 540 in total. The data collection flow is illustrated in Figure 4. The median time taken by participants to complete the study was approximately 12.26 minutes, and the average time was slightly higher at 14.58 minutes, with a standard deviation of 7.85 minutes. Initially, each participant received a compensation of two US dollars for their time. However, for the groups where the average time spent exceeded 12 minutes, we increased the incentives according to the rule of Prolific. This adjustment ensured that the average hourly reward was maintained above ten US dollars.

We collected different types of user behavioral data (explained below) from participants to answer the RQs:

Clicking behavior: We collected data on the number of clicks on each SERP. Only the initial click on an article was taken into account. Each click on a search result triggered the display of the full text of the clicked result. Related contextual data, such as the

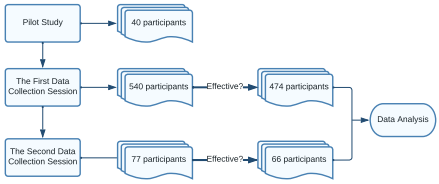


Figure 4: Data Collection Workflow.

study group, sequence of the SERP in the corresponding session, and clicking timestamp, were also automatically recorded. With the timestamp, we were able to identify the clicking behavior of a participant before they clicked a target result on the same SERP.

Dwell Time: We recorded the timestamps each time a participant opened and subsequently closed the full text of a search result. These two timestamps were utilized to calculate the length of time (referred to as dwell time) that the participant spent reading the full text of each clicked search result.

Assessment Data: After closing the full text of a clicked result, participants were required to evaluate the credibility and usefulness levels of the search result in relation to the task question (See Step 6 in Figure 4).

With the data collected at different stages of search interactions, we did a comprehensive comparison between the control and treatment groups. We also took into account three major contextual factors that may have an impact on decoy effects in the analysis. Firstly, the prior knowledge could affect their credibility and usefulness judgments. Secondly, the size of decoy effect in search may depend on on whether people perceived the salient misinformation from the decoy results. Whether participants have read the full text of a decoy result could also influence their in-situ vulnerability to decoy items and their subsequent evaluations of target results. Therefore, we analyzed the behavioral impact of clicks on decoy results prior to target results. Another measure to examine if a participant perceive decoy results as inferior option is their evaluation. For this, we included their ratings on decoy results in our analysis.

Furthermore, to better address RQ2, we also incorporated a set of contextual factors into regression analysis to obtain a better understanding of the moderating effects of these factors on decoy effect in search and judgments. These factors include participants' behavior, their ratings of *non-target* results, presentation features of target results, and participants' self-reported prior medical knowledge on COVID treatments. We chose Ordinal Logistic Regression for the analysis because our dependent variable is discrete ordered ratings given by participants, which range from one to five. We employed regression analysis on credibility and usefulness judgments respectively.

5 Results

This section presents the results of our analysis in response to the proposed research questions.

5.1 Descriptive Statistics

As a prepration for addressing the RQs, we presented the descriptive statistics of users' search behaviors in this section. In particular,

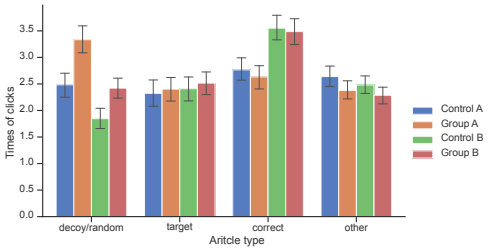


Figure 5: Average times of clicking different types of articles per participant.

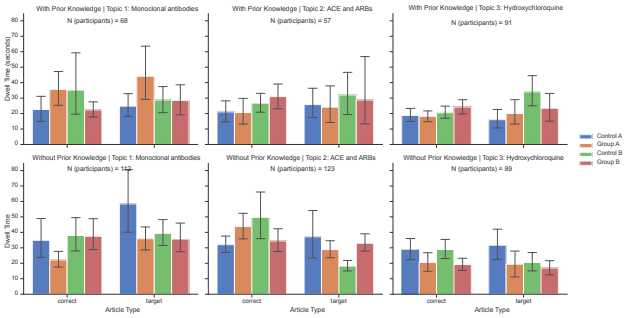


Figure 6: Average dwell time participants spent on articles.

the presented behavioral features will be useful for understanding the moderating effects of contextual factors on decoy effects in judgments (i.e. RQ2).

5.1.1 Click behavior. Figure 5 shows the average number of clicks on different types of articles. Regarding rank position bias, we found that result ranking plays a dominant role in clicks. For example, correct articles were clicked more in Group B and Control B than Group A and Control A as the correct article is the first item in the ranked result list in B conditions but the third item in A conditions. Another key factor is the level of *relevance*. Even decoy articles are salient misinformation but they have higher level of relevance compared to the random results controlled as part of the "background" for experimental conditions. As a result, decoy articles were clicked more frequently than random ones.

The target articles were clicked slightly more in treatment groups where a decoy article was presented in the result list, compared to control groups. However, the difference is not statistically significant. This indicates that decoy items generate less impact on clicking compared to rank positions and topical relevance, and that we should examine the behavioral impact of decoy on other dimensions.

5.1.2 Dwell Time. Since prior knowledge of the topic may affect participants' dwell time in search [27, 45], we analyzed the dwell time of participants with and without prior knowledge separately. We excluded records with less than 3 seconds and two records with more than 1,000 seconds. Figure 6 shows the average time participants spent on correct and target articles under different conditions. No consistent pattern was observed in terms of dwell time on correct results. For target articles, we found a consistent between-group difference pattern in dwell time of participants without prior knowledge. Specifically, participants without prior knowledge in Group A were more likely to spend less time on target articles than participants without prior knowledge in Control A. The difference in dwell time on target articles between Group A and Control A is significant in Topic 1 and Topic 3. For Topic 1, participants from Group A spent 29 seconds less than participants from Control A in reading a target article on average ($t = -2.245, p = 0.015$). For Topic 2, participants from Group A spent 15.9 seconds less than participants from Control A on average ($t = -1.842, p = 0.068$). For Topic 3, participants from Group A spent 10.9 seconds less than participants from Control A on average ($t = -1.302, p = 0.013$). This indicates that the decoy articles in Group A could encourage participants without prior knowledge to spend less time reading target articles. However, no similar pattern was found between Group B and Control B, which may be because the size of decoy effect is smaller in Group B than in Group A.

5.2 Credibility and Usability Assessments

RQ1 focuses on how decoy search results, operationalized by pages with low levels of credibility and usefulness, affect participants' usefulness and credibility judgments, especially on target medical misinformation results.

Before investigating decoy effect, we first perform a *treatment check* to verify that participants in each of the four groups perceived the intended levels of credibility and usefulness on three types of search results (i.e. decoy, correct, and target results). Participants' mean (standard deviation) credibility scores were 3.36 (1.22), 3.87 (0.95), and 3.79 (0.94) for decoy, correct, and target results, respectively; Participants' mean (standard deviation) were 3.12 (1.26), 3.82 (1.06), and 3.63 (1.10) for decoy, correct, and target results. This result suggests that participants did experience the intended treatments (i.e. assigning higher scores to the target misinformation result than to decoy misinformation) for both credibility and usefulness judgments. The decoy results were expected to appear with lower levels of credibility and usefulness, so the decoy results would serve as a *reference point* that makes the target results more preferable. A bar chart presenting these average credibility and usefulness scores assigned by participants in each group is in Figure 7. The solid line error bars represent standard deviations. The mean scores of these results are consistent with the expected decoy impacts suggested in previous behavioral experiments [32, 73].

The difference between the mean scores of target and correct results were smaller under treatment conditions than that in control conditions (see Figure 7). We conducted t-tests to investigate the statistical significance of this possible decoy effects. The results indicates that the difference between average credibility scores assigned to target results by participants in the treatment and control group did not differ significantly ($t = 0.83, p = 0.20$). The difference between average usefulness scores in the treatment and control group was not significant either ($t = 1.05, p = 0.15$). This may be because our user study was restricted by the specific topic (i.e. COVID treatment misinformation) or possible mixed contextual

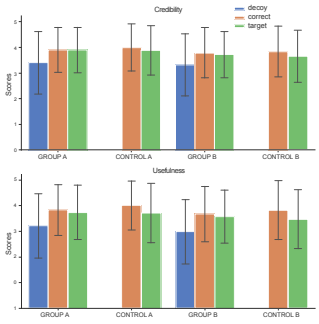


Figure 7: Distribution of mean credibility and usefulness scores under four conditions.

impacts. It is also possible that decoy effects may not be effective without certain behaviors as prior conditions. For example, if a participant did not read the title and SERP, or skip full text of a decoy result, the decoy result would not affect the participant's judgment of other results, including target and correct results. In the following subsection, we explore how various interactions of participants with decoy results and other contextual factors affect their credibility and usefulness judgment.

5.3 Moderating Effect of Search Contextual Factors

To answer RQ2, this subsection reports the result on the moderating effects of search contextual factors, which can better reveal the conditions under which decoy items generate significant behavioral impacts. Specifically, we conducted further analysis with *subgroups* defined based on prior knowledge, clicking behavior and decoy result ratings. For instance, participants' clicking on decoy results may indicate a deeper interaction and trust on the results (compared to merely browsing them on SERPs). As a result, it may affect the decoy effect on users' following click behaviors, dwell time on pages (especially target results), and usefulness and credibility ratings. Adding these binned analyses can enhance our understanding of the conditions and contexts under which decoy options affect individuals the most. The results are explained below.

5.3.1 Impact of prior knowledge and decoy clicking behavior. Before exploring the impact of decoy clicking behavior, we examined the role of *prior knowledge* on the search process, as previous research suggested that prior knowledge may affect participants' decisions on what results to click or skip [e.g. 55, 65]. We asked participants to answer the same binary question that they would answer in their later search task, but they were allowed to answer "Not sure" if they don't have prior knowledge for answering the question. 62.2% of participants (112 out of 180) in Topic 1, 68.3% of participants (123 out of 180) in Topic 2, and 50% of participants in Topic 3 did not have prior knowledge about the topic.

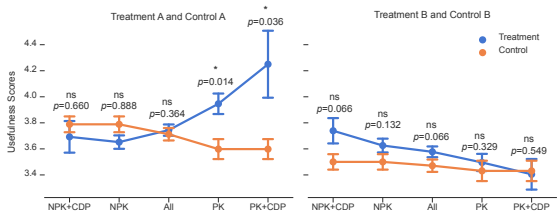


Figure 8: Distribution of mean usefulness scores between treatment and control groups under five conditions. All: all participants; PK: participants with prior knowledge; NPK: participants without prior knowledge; CDP: participants who clicked decoy results.

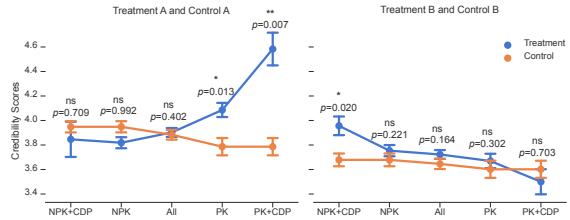


Figure 9: Distribution of mean credibility scores in treatment and control groups under five conditions.

Since prior knowledge may affect search behavior and judgment [60, 69], we explored the impact of prior knowledge, moderated by decoy-clicking behavior. Figure 8 and 9 present the distributions of mean usefulness and mean credibility scores assigned to target results, with 95% confidence intervals. These results are grouped by two variables: whether participants engaged in decoy-clicking behavior and their level of prior knowledge. The decoy clicking variable measures if a participant clicked the decoy result presented on the same SERP with the target result that the participant rated its credibility and usefulness. The patterns are similar between credibility and usefulness scores. We presented the differences in participants' ratings of credibility and usefulness between control and treatment groups under different conditions. The conditions were grouped by participants' prior knowledge (PK vs. NPK) and whether participants clicked on the decoy result before they clicked on the target result (CDP). The blue line represents the mean scores assigned by participants from Treatment groups, while the orange line represents the mean scores assigned by participants from Control groups. Using t-tests, we assessed the statistical significance of differences between these groups under varying conditions. We found that participants with prior knowledge (PK) in Treatment A consistently assigned higher usefulness and credibility scores to target results than those in Control A. This effect was more significant when participants clicked on the decoy result prior to evaluating the target result. Clicking on a decoy result before the target appeared to amplify the decoy's influence, particularly in Treatment A, where decoy results were ranked higher than the target misinformation results.

The reason why participants with prior knowledge were more likely to be affected by decoy effects might be because they have

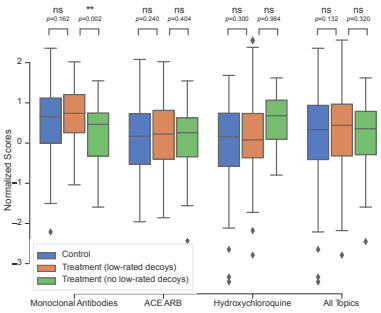


Figure 10: Comparison of Usefulness Scores: Control vs. Treatment Groups. The treatment groups are distinguished based on whether participants assigned low usefulness ratings to the decoy results.

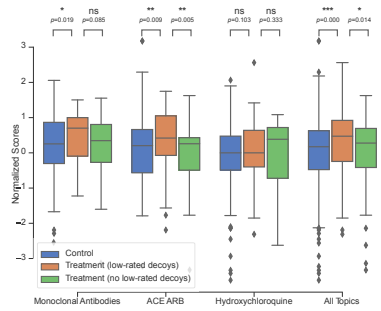


Figure 11: Comparison of Credibility Scores: Control vs. Treatment Groups. The treatment groups are distinguished based on whether participants assigned low credibility ratings to the decoy results.

better judgments on credibility and domain-specific document usefulness. They could perceive the low credibility and limited usefulness of decoy items more accurately than participants without prior knowledge, which makes target results more preferable in comparison. Since the decoy impacted the judgment of participants with prior knowledge who did not click decoy results, they probably perceived the low quality of decoy results due to the lower-quality titles and snippets with limited specific information. However, there was no significant difference between scores assigned to targets in Treatment B and Control B, indicating that the rank position of decoy results on SERPs could play an important role in the decoy effects.

5.3.2 *Impact of credibility and usefulness ratings on decoy results.*

The participants' judgment on decoy results could affect the size of decoy effects. If participants did not perceive the decoy result as a reference point, then the decoy may have had little to no impact on their judgment accuracy on target misinformation. To explore this, we investigated the impact of low credibility and usefulness scores assigned to decoy results. *Normalized* scores were used in this analysis to avoid the leniency and strictness bias of participants. The normalized scores were calculated by $[s - mean(s_i)]/std(s_i)$, in which s is the specific score to be normalized, s_i is the set of scores assigned to search results by the same participant who provided s. A normalized score assigned to a result is less than zero indicating the result was rated lower than the participant's avareage, classifying it as a low-rated score.

Figure 10 and Figure 11 show a comparison of scores assigned to the target results across three groups of participants: Control Group, in which participants did not interact with a decoy result; Low-Rating Group, in which participants assigned a low-rated score (normalized score < 0) to a decoy result displayed on the same page as the target result; and Non-Low Rating Group, in which participants assigned a non-low rating (normalized score ≥ 0) to a decoy result on the same page as the target result. This comparison allows us to analyze how participants' evaluations of decoy results, particularly low ratings, influenced their judgments of target results.

The credibility scores assigned to target results by participants who assigned a low score to the decoy result on the same SERP are significantly higher than those of participants in a control group and participants who did not assign a low score to the decoy. This is evident for the topics *Monoclonal Antibodies* and *ACE / ARB*, with significant p-values of 0.019 and 0.009. This suggests that participants who recognized decoys as low-quality were more favorable toward the credibility of target results. However, this effect did not extend to perceived usefulness. For the same topics (Monoclonal Antibodies and ACE/ARB), the differences in usefulness scores between groups were not significant. This indicates a divergence between how participants judged the credibility of information and how they rated its usefulness.

Table 2: A mixed model analysis of usefulness scores assigned to target results.

	Treatment Group		Control Group	
IV	Estimate	Std. Error	Estimate	Std. Error
ClickedDecoyFirst	0.399*	0.179		
ClickedRandomFirst			0.127	0.231
TimeReadDecoy	-0.006*	0.002		
TimeReadRandom			0.0004	0.006
TimeReadTarget	0.008**	0.002	0.003*	0.001
DecoyAhead	0.675**	0.252		
RandomAhead			0.545·	0.276
Page	0.202***	0.068	0.287**	0.076
PriorKnowledge	0.371	0.261	-0.114	0.277
AIC	2508.4		2196.0	
nobs	948		813	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '·'

5.4 Regression Analysis

A mixed model analysis was performed to further reveal the impact of contextual factors under different conditions (RQ2). This model incorporates both fixed effects, represented by independent variables, and random effects of participants and topics, which account for random variability at the participant and topic levels. The analysis focuses on two dependent variables: the credibility scores and usefulness scores that users assigned to target results, as detailed

in Table 2 and Table 3, respectively. A comprehensive analysis was conducted on multiple independent variables to further understand their impacts on users' context-dependent preferences.

Table 3: A mixed model analysis of credibility scores assigned to target results.

	Treatment Group		Control Group	
	Estimate	Std. Error	Estimate	Std. Error
ClickedDecoyFirst	0.522*	0.207		
ClickedRandomFirst			0.082	0.251
TimeReadDecoy	-0.001	0.003		
TimeReadRandom			0.014	0.007
TimeReadTaget	0.006*	0.003	0.002	0.001
DecoyAhead	0.875**	0.323		
RandomAhead			0.522	0.346
Page	0.161*	0.075	0.070	0.082
PriorKnowledge	0.413	0.335	-0.339	0.352
AIC	2079.6		1889.2	
nobs	948		813	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

We selected independent variables that may influence participants' perceptions of credibility and usefulness based on the analyses reported above and evidence from previous literature [e.g. 49, 82]. We categorized the independent variables into three groups. Initially, we selected variables that may be associated with significant behavioral differences. If a participant clicked on a decoy result before clicking on a target result, or if a participant spent more time reading the full text of a decoy result on the same SERP, the decoy effect induced by the decoy result might be more potent than if the decoy result was clicked after the target result. Therefore, we have included a variable, '**ClickedDecoyFirst**', to denote whether a participant clicked on the decoy result (treatment). For the control group, correspondingly, '**ClickedRandomFirst**' denotes clicking on the random result (control) prior to clicking the target result. We have also included '**TimeReadDecoy**' to measure the time (in seconds) a participant spent reading a decoy result in a treatment group and '**TimeReadRandom**' to measure the time a participant spent reading a random result in a control group. Additionally, dwell time on the target result could also influence their judgment of the misinformation presented by the result. Hence, we have included another variable, '**TimeReadTarget**', to measure this time (in seconds).

We also included a set of variables to characterize the presentation features of target results that may also affect participants' credibility and usefulness judgments. '**DecoyAhead**' describes if a decoy result in a treatment group is ranked higher than the target result on the same page, and '**RandomAhead**' describes if a random result is ranked higher than the target result in a control group. Another variable, '**Page**', is the sequence number (1-4) of the query/SERP where the target result is presented. Finally, the participants' prior knowledge (**PriorKnowledge**) about the topic they searched is also included in the regression analysis.

Clicking a decoy result prior to clicking a target result (**ClickDecoyFirst**) has positive effects on the credibility and usefulness scores assigned to the target misinformation result, indicating that participants might be more vulnerable to decoy effects if they read a decoy article on the same SERP before reading the target article. However, reading a random-topic article prior to reading a target article **(ClickRandomFirst)** does not affect the credibility and usefulness scores assigned to the target result. This means that reading an irrelevant result may not affect participants' judgments on articles they read afterward. Compared to the control group, we observed significant decoy effects that led to higher scores on target articles when participants clicked a decoy result first. This finding is consistent with the analysis in Section 5.3.1 in which we found that clicking a decoy article prior to clicking a target could amplify decoy effects.

We also found that dwell time on a decoy result (**TimeReadDecoy**) had significant negative effects on the usefulness scores assigned to the target article on the same SERP. However, a similar effect was not observed on credibility judgments. Dwell time on random-topic articles (**TimeReadRandom**) had no significant effect on the credibility evaluation of target articles. In addition, we investigated if the dwell time on target results (**TimeReadTarget**) would affect outcomes. Participants tended to assign significantly higher usefulness scores to target articles in both the treatment and control groups, but the effect of dwell time on usefulness scores was stronger in the control group. In regard to credibility scores, the dwell time on target results (**TimeReadTarget**) was found to have a significant positive effect on credibility scores in the treatment group, but this effect was not significant in the control group. These findings could imply that the presence of decoy effects in treatment conditions may enhance the perceived usefulness and credibility of the target article when more time is spent reading it, more than the control conditions.

With respect to the interaction effect between decoy and rank position bias, we found that if a decoy webpage is ranked above a target result, the decoy effect on participants' evaluation of target results tended to be stronger. The variable **DecoyAhead** refers to the scenarios where a decoy article is ranked above a target. Our result indicates that this variable has positive effects on the usefulness and credibility scores assigned to the target article. The counterpart, **RandomAhead**, indicating a random article is ranked above a target, has no significant effect on usefulness and credibility. **Page**, which indicates the query/SERP sequence in a session, had a significant positive effect on both usefulness and credibility scores in the treatment group, suggesting that decoy effects tend to be stronger in later query segments or SERPs of a session. When the decoy item is removed (i.e. control group), similar query sequence effect was not observed. Regarding **PriorKnowledge**, the results indicate that participants' pre-existing knowledge about the topic had no significant effect on their assessments of usefulness and credibility. **PriorKnowledge** had a positive effect on both scores in treatment conditions, while it had a negative effect on both scores in the control condition.

The analysis reveals that clicking on a decoy result before a target article increases the perceived usefulness and credibility of the target, suggesting a priming effect of decoys. Dwell time on the decoy may not enhance the perceived usefulness of the target, but not its credibility, while more time spent on the target article itself increases its usefulness and credibility in the treatment group. The position of a decoy above the target amplifies this effect, indicating

that article ranking influences user perceptions. Participants' prior knowledge does not significantly affect their evaluations.

6 Discussion

In contrast to the assumptions implicitly made in a variety of formal models, users are *boundedly rational* and usually do not make search decisions based on accurately estimated search gains and costs [5, 16, 42, 78]. With the mainstream computing research focusing on data and algorithmic biases [37], it is critical to pay attention to, capture, and properly address the negative impact of biases from the human side [43]. Among diverse cognitive biases that affect individuals' decisions, *decoy effect* offers a more realistic explanation compared to simplified rational models on how and why users alter their in-situ preferences and judgments on presented options and has been empirically confirmed by a wide range of behavioral experiments [73]. However, it is still unclear how decoy search results affect users' credibility and usefulness judgments in session contexts and how this decoy effect interact with search contextual factors, such as topical knowledge, rank position, and prior search experience. In particular, it is essential to investigate how users react to online misinformation that involves high-stakes medical decisions under the impact of decoy items.

To address this gap and take a step forward towards the vision of human-centered intelligent IR, our study examined the role of decoy effect and the behavioral conditions triggering it in COVID medical information evaluation with a between-subject crowdsourcing experiment. Based on the results collected from 540 participants (2,160 valid SERP-based interactions) under three topics and four conditions, we have following answers to the research questions:

Regarding **RQ1**, our results show that users are more vulnerable to the negative effect of decoy results in usefulness and credibility judgments and click on decoy results in a higher rate when they have certain level of prior knowledge on the topic. Regarding rank position, results from treatment condition **B** suggest that decoy effect may not occur when the decoy option is ranked below the other results being evaluated. In addition, when user explicitly recognizes the decoy result as a low-quality result, it could further increase the risk of falling into the pitfall of COVID misinformation, especially under the subtopics that most people are not familiar with or do not have a strong prior opinion on (i.e. Monoclonal Antibodies and ACE/ARB). Differing from the observations from classical/simplified behavioral experiments [e.g. 8, 75, 83], our findings indicate that in Web search sessions, a potential decoy result presented on SERP does not necessarily lead to significant decoy effects. Although users' prior knowledge may facilitate the development of search tactics and evaluation criteria, it may also increase the tendency to relying on familiar signals (e.g. names of entities and institutions, statistics) which lead to a false perception of high credibility levels and higher risk of decoy effects. In addition, richer interactions with decoy results (e.g. clicking and assigning a low rating) can strengthen the decoy effect on users' context-dependent judgments. Going beyond document assessment experiments that abstract out search factors [e.g. 3, 25, 67], our findings were collected from session-based interaction settings, identified potential triggers of decoy effects, and revealed the interplay of decoy options, user characteristics, and rank positions.

With respect to **RQ2**, the results from regression modeling demonstrate that longer dwell time on and higher ranking of decoy results tend to cause higher credibility ratings on COVID treatment misinformation. Significant decoy effects were also observed when a decoy result was clicked first before the target misinformation was examined by a user. Apart from users' interactions with decoy results, we found that a user's credibility rating on COVID misinformation tended to be higher when the user possessed certain level of prior knowledge and viewed more SERPs. These decoy-related effects are not significant under the topic of Hydroxychloroquine. This may be because Hydroxychloroquine previously received more attention through social media platforms and national media coverage than the other two candidate treatments [4]. As a result, many users may already have a prior established belief on this topic. Similar patterns of decoy effects were also observed in usefulness ratings. Findings under RQ2 go beyond traditional document-assessment-only experiments and clarify the behavioral and session-level factors that may moderate decoy effects.

Inspired by behavioral economics approach to modeling users, our study thoroughly examined decoy effect in users' information judgments, especially their in-situ vulnerability to COVID treatment misinformation on SERPs. Findings from our experiment characterized the role and variation of decoy effect under diverse rank positions, levels of user topical knowledge, and subtopics in simulated search sessions and could inform the training of bias prediction models and the development of bias-aware interactive search systems and evaluation metrics. Our study goes beyond traditional offline evaluation experimental setup [e.g. 3, 25, 64, 67] and examined users' context-dependent preferences and their connections to decoy effect in simulated session contexts with predefined SERPs. Also, with the knowledge about decoy effect, researchers and system designers can design bias-aware recommendations, interventions, and intelligent nudging techniques for protecting users (especially the ones who are particularly vulnerable to certain biases under the moment) from health misinformation [31]. More broadly, our research indicates the value of representing decoy effect in user models and incorporating the knowledge into the prediction of user judgments. With the experimental design and empirical evidences from our work, IR researchers can further explore other forms of cognitive biases introduced by behavioral economics research [e.g. 32, 42, 73] and examine their behavioral effects. Recognizing the *bounded rationality* of users and characterizing their cognitive biases in Web search will enable researchers to integrate the insights from behavioral economics with algorithms and evaluation metrics, develop scalable psychology-informed assistance tools for combating misinformation, and build a behaviorally more realistic foundation for next-generation search systems.

Similar to other works, our study has limitations as well as implications for future studies in this interdisciplinary field. Our design of decoy option, which was restricted to search result snippets only, may not fully trigger or reflect potential decoy effects hidden in real-life search activities. However, this deliberate design decision allowed us to better balance the goal for measuring decoy effect under a controlled environment and the need for enhancing the authenticity of simulated search experience. Restricting the decoy result design within the snippets reduced potential mixed effects from SERP components and content pages and thereby helped us

better clarify the conditions and behavioral triggers of decoy effect. Our findings on SERP-level decoy effects will inform the design and meta-evaluation of ad hoc offline evaluation metrics, especially in terms of estimating click probabilities and rank-position-based utility decay in ranked result lists [cf. 16, 42, 54]. Future research can also expand the design of decoy conditions and further investigate the system and contextual triggers of decoy effect at different search decision points and in different domains [15].

Also, due to the limitation of the crowdsourcing platform and concerns about data quality, we could not deploy fully interactive, long-session experiments and naturalistic work tasks. Instead, we implemented predefined SERPs and sessions to observe and compare user behaviors and judgments across different conditions. It is possible that the (short) length of sessions and controlled interaction patterns may reduce the variations in user actions and restrict users' search strategies, which lead to the difficulty of observing potentially significant effects of decoy options. In future studies, researchers can further explore a diverse set of decoy items in different search modalities, and investigate how decoy results affect users' perceptions, engagement, and search evaluations.

Another main limitation is that our experiment involves the judgments of COVID treatment information only. It is unclear how the size and triggers of decoy effect would vary under diverse domains and modalities of interactions (e.g. traditional SERP-based search versus AI-enabled chat search). In addition, our investigation on participants' vulnerability to decoy is mainly based on their *in-situ* behaviors (e.g. clicks and dwell time on target misinformation and decoy results) and judgments (e.g. credibility and usefulness rating on target results). Although participants' answers to the question is not our main focus for decoy effect examination, it is possible that the participants might guess the answer of binary questions correctly (e.g. due to prior beliefs or by chance). For future work, researchers may deploy a more open-ended approach to capture prior knowledge information and add a pre-interaction test session to reliably measure the levels of participants' knowledge on the topic being studied in a lab or naturalistic study setting. Nevertheless, COVID treatment information has been a highly relevant topic to different populations during the COVID pandemic, and how to combat rapidly spreading misinformation on this topic and improve the informational wellbeing of the general public (especially marginalized communities) remains a global challenge both within and beyond research communities. Investigating decoy effect on users' vulnerability to medical misinformation can produce practically useful knowledge for addressing this challenge to public health and help tackle the negative impacts of other biases.

7 Conclusion

Examining and addressing biases is essential for building fair, inclusive information ecosystems [43]. However, investigating data and algorithmic bias alone cannot offer us the full picture. Characterizing and modeling human biases and bounded rationality encourages us to re-examine the foundations of existing user models and behavioral assumptions and can take us further towards building truly person-centered intelligent information access systems and contribute to a balanced fairness approach that considers biases from both human and system sides [43]. Our study focused on context-dependent decoy effects in medical misinformation judgment and revealed how decoy effects vary under different user characteristics, search interactions, SERP conditions and rank positions. This *bias-aware* viewpoint can also be applied in enhancing other forms of human-information interaction and human-AI interaction, such as users' interactions with and judgment of contents from large language models (LLMs) and generative search engines [14, 47, 72], as well as in assessing *systems' vulnerability* to decoy bias when retrieving and presenting search results [15]. Our experiment on decoy effect can serve as an initial step towards achieving this vision.

References

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

[2] Azzah Al-Maskari and Mark Sanderson. 2010. A review of factors influencing user satisfaction in information retrieval. *Journal of the American Society for Information Science and Technology* 61, 5 (2010), 859–868.

[3] Omar Alonso and Stefano Mizzaro. 2012. Using crowdsourcing for TREC relevance assessment. *Information processing & management* 48, 6 (2012), 1053–1066.

[4] Hubert Au, Jonathan Bright, and Philip N Howard. 2020. Social Media Junk News on Hydroxychloroquine and Trust in Science. *Coronavirus Misinformation Weekly Briefing 03-08-2020* (2020).

[5] Leif Azzopardi. 2021. Cognitive biases in search: a review and reflection of cognitive biases in Information Retrieval. In *Proceedings of the 2021 conference on human information interaction and retrieval*. 27–37.

[6] Leif Azzopardi and Jiqun Liu. 2024. Search under Uncertainty: Cognitive Biases and Heuristics-Tutorial on Modeling Search Interaction using Behavioral Economics. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval*. 427–430.

[7] Ricardo Baeza-Yates. 2020. Bias in search and recommender systems. In *Fourteenth ACM Conference on Recommender Systems*. 2–2.

[8] Ian J Bateman, Alistair Munro, and Gregory L Poe. 2008. Decoy effects in choice experiments and contingent valuation: Asymmetric dominance. *Land Economics* 84, 1 (2008), 115–127.

[9] Nicholas J Belkin. 2016. People, interacting with information. In *ACM SIGIR Forum*, Vol. 49. ACM New York, NY, USA, 13–27.

[10] Nolwenn Bernard and Krisztian Balog. 2023. A Systematic Review of Fairness, Accountability, Transparency and Ethics in Information Retrieval. *Comput. Surveys* (2023).

[11] Jennifer S Blumenthal-Barby and Heather Krieger. 2015. Cognitive biases and heuristics in medical decision making: a critical review using a systematic search strategy. *Medical Decision Making* 35, 4 (2015), 539–557.

[12] Nattapat Boonprakong, Xiuge Chen, Catherine Davey, Benjamin Tag, and Tilman Dingler. 2023. Bias-Aware Systems: Exploring Indicators for the Occurrences of Cognitive Biases when Facing Different Opinions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[13] Ben Carterette, Paul Clough, Mark Hall, Evangelos Kanoulas, and Mark Sanderson. 2016. Evaluating retrieval over sessions: The TREC session track 2011-2014. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 685–688.

[14] Nuo Chen, Jiqun Liu, Xiaoyu Dong, Qijiong Liu, Tetsuya Sakai, and Xiao-Ming Wu. 2024. AI Can Be Cognitively Biased: An Exploratory Study on Threshold Priming in LLM-Based Batch Relevance Assessment. *arXiv preprint arXiv:2409.16022* (2024).

[15] Nuo Chen, Jiqun Liu, Hanpei Fang, Yuankai Luo, Tetsuya Sakai, and Xiao-Ming Wu. 2024. Decoy Effect In Search Interaction: Understanding User Behavior and Measuring System Vulnerability. *arXiv preprint arXiv:2403.18462* (2024).

[16] Nuo Chen, Jiqun Liu, and Tetsuya Sakai. 2023. 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*. 3396–3405.

[17] Nuo Chen, Jiqun Liu, Tetsuya Sakai, and Xiao-Ming Wu. 2023. Decoy Effect in Search Interaction: A Pilot Study. *arXiv preprint arXiv:2311.02362* (2023).

[18] Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. 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*. 15–24.

[19] Michael Cole, Jingjing Liu, Nicholas J Belkin, Ralf Bierig, Jacek Gwizdka, Chang Liu, Jin Zhang, and Xiangmin Zhang. 2009. Usefulness as the criterion for evaluation of interactive information retrieval. In *Proceedings of the Third Workshop on Human-Computer Interaction and Information Retrieval Cambridge*. HCIR, 1–4.

[20] Terry Connolly, Jochen Reb, and Edgar E Kausel. 2013. Regret salience and accountability in the decoy effect. *Judgment and Decision making* 8, 2 (2013), 136–149.

[21] Jose Yunam Cuan-Baltazar, Maria José Muñoz-Perez, Carolina Robledo-Vega, Maria Fernanda Pérez-Zepeda, and Elena Soto-Vega. 2020. Misinformation of COVID-19 on the internet: infodemiology study. *JMIR public health and surveillance* 6, 2 (2020), e18444.

[22] Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. 2023. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in Public Health* 11 (2023), 1166120.

[23] Nicola Diviani, Bas van den Putte, Stefano Giani, and Julia CM van Weert. 2015. Low health literacy and evaluation of online health information: a systematic review of the literature. *Journal of medical Internet research* 17, 5 (2015), e112.

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

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

[26] Donna Harman. 2011. *Information retrieval evaluation*. Morgan & Claypool Publishers.

[27] Daniel Hienert, Matthew Mitsui, Philipp Mayr, Chirag Shah, and Nicholas J Belkin. 2018. The role of the task topic in web search of different task types. In *Proceedings of the 2018 conference on human information interaction & retrieval*. 72–81.

[28] Brian Hilligoss and Soo Young Rieh. 2008. Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing & Management* 44, 4 (2008), 1467–1484.

[29] Jianping Hu and Rongjun Yu. 2014. The neural correlates of the decoy effect in decisions. *Frontiers in behavioral neuroscience* 8 (2014), 271.

[30] Joel Huber, John W Payne, and Christopher Puto. 1982. Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of consumer research* 9, 1 (1982), 90–98.

[31] Dennis Hummel and Alexander Maedche. 2019. How effective is nudging? A quantitative review on the effect sizes and limits of empirical nudging studies. *Journal of Behavioral and Experimental Economics* 80 (2019), 47–58.

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

[33] Daniel Kahneman, Stewart Paul Slovic, Paul Slovic, and Amos Tversky. 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.

[34] Markus Kattenbeck and David Elsweiler. 2019. Understanding credibility judgements for web search snippets. *Aslib Journal of Information Management* (2019).

[35] Diane Kelly and Leif Azzopardi. 2015. How many results per page? A study of SERP size, search behavior and user experience. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 183–192.

[36] Youngwoo Kim, Razieh Rahimi, and James Allan. 2022. Alignment Rationale for Query-Document Relevance. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2489–2494.

[37] Nima Kordzadeh and Maryam Ghasemaghaei. 2022. Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems* 31, 3 (2022), 388–409.

[38] Mucahid Kutlu, Tyler McDonnell, Tamer Elsayed, and Matthew Lease. 2020. Annotator rationales for labeling tasks in crowdsourcing. *Journal of Artificial Intelligence Research* 69 (2020), 143–189.

[39] Sook Lim. 2013. College students' credibility judgments and heuristics concerning Wikipedia. *Information Processing & Management* 49, 2 (2013), 405–419.

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

[41] Jiqun Liu. 2022. Toward Cranfield-inspired reusability assessment in interactive information retrieval evaluation. *Information Processing & Management* 59, 5 (2022), 103007.

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

[43] Jiqun Liu. 2023. Toward A Two-Sided Fairness Framework in Search and Recommendation. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. 236–246.

[44] Jiqun Liu and Leif Azzopardi. 2024. 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*. 3013–3016.

[45] Jingjing Liu, Michael J Cole, Chang Liu, Ralf Bierig, Jacek Gwizdka, Nicholas J Belkin, Jun Zhang, and Xiangmin Zhang. 2010. Search behaviors in different task types. In *Proceedings of the 10th annual joint conference on Digital libraries*. 69–78.

[46] Jiqun Liu and Fangyuan Han. 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*. 1141–1150.

[47] Jiqun Liu and Jiangen He. 2024. The Decoy Dilemma in Online Medical Information Evaluation: A Comparative Study of Credibility Assessments by LLM and Human Judges. *arXiv preprint arXiv:2411.15396* (2024).

[48] Jiqun Liu and Chirag Shah. 2019. *Interactive IR user study design, evaluation, and reporting*. Morgan & Claypool Publishers.

[49] Jiqun Liu and Chirag Shah. 2019. Investigating the impacts of expectation disconfirmation on web search. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. 319–323.

[50] Jiqun Liu and Chirag Shah. 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*. 1–11.

[51] Jiaying Liu, Yan Zhang, and Yeolib Kim. 2023. Consumer health information quality, credibility, and trust: an analysis of definitions, measures, and conceptual dimensions. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. 197–210.

[52] Irene Lopatovska. 2014. Toward a model of emotions and mood in the online information search process. *Journal of the association for information science and technology* 65, 9 (2014), 1775–1793.

[53] N Gregory Mankiw. 2014. *Principles of economics*. Cengage Learning.

[54] Alistair Moffat, Joel Mackenzie, Paul Thomas, and Leif Azzopardi. 2022. A flexible framework for offline effectiveness metrics. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 578–587.

[55] Sophie Monchaux, Franck Amadieu, Aline Chevalier, and Claudette Mariné. 2015. Query strategies during information searching: Effects of prior domain knowledge and complexity of the information problems to be solved. *Information Processing & Management* 51, 5 (2015), 557–569.

[56] Heather L O'brien and Elaine G Toms. 2013. Examining the generalizability of the User Engagement Scale (UES) in exploratory search. *Information Processing & Management* 49, 5 (2013), 1092–1107.

[57] Gordon Pennycook and David G Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116, 7 (2019), 2521–2526.

[58] Daniela Petrelli. 2008. On the role of user-centred evaluation in the advancement of interactive information retrieval. *Information Processing & Management* 44, 1 (2008), 22–38.

[59] Jonathan C Pettibone and Douglas H Wedell. 2000. Examining models of nondominated decoy effects across judgment and choice. *Organizational behavior and human decision processes* 81, 2 (2000), 300–328.

[60] Zuzana Pinkosova, William J McGeown, and Yashar Moshfeghi. 2023. Moderating effects of self-perceived knowledge in a relevance assessment task: An EEG study. *Computers in Human Behavior Reports* 11 (2023), 100295.

[61] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*. 1003–1012.

[62] Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2021. Vera: Prediction techniques for reducing harmful misinformation in consumer health search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2066–2070.

[63] Kevin Roitero, Alessandro Checco, Stefano Mizzaro, and Gianluca Demartini. 2022. Preferences on a budget: Prioritizing document pairs when crowdsourcing relevance judgments. In *Proceedings of the ACM Web Conference 2022*. 319–327.

[64] Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Falk Scholer. 2021. On the effect of relevance scales in crowdsourcing relevance assessments for Information Retrieval evaluation. *Information Processing & Management* 58, 6 (2021), 102688.

[65] Mylene Sanchiz, Aline Chevalier, and Franck Amadieu. 2017. How do older and young adults start searching for information? Impact of age, domain knowledge and problem complexity on the different steps of information searching. *Computers in Human Behavior* 72 (2017), 67–78.

[66] Laura Sbaffi and Jennifer Rowley. 2017. Trust and credibility in web-based health information: a review and agenda for future research. *Journal of medical Internet research* 19, 6 (2017), e218.

[67] Falk Scholer, Diane Kelly, Wan-Ching Wu, Hanseul S Lee, and William Webber. 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*. 623–632.

[68] Itamar Simonson. 1989. Choice based on reasons: The case of attraction and compromise effects. *Journal of consumer research* 16, 2 (1989), 158–174.

[69] Catherine L Smith and Soo Young Rieh. 2019. Knowledge-context in search systems: toward information-literate actions. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. 55–62.
[70] Michael Soprano, Kevin Roitero, David La Barbera, Davide Ceolin, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2021. The many dimensions of truthfulness: Crowdsourcing misinformation assessments on a multidimensional scale. *Information Processing & Management* 58, 6 (2021), 102710.
[71] Sandro Tiziano Stoffel, Jiahong Yang, Ivo Vlaev, and Christian von Wagner. 2019. Testing the decoy effect to increase interest in colorectal cancer screening. *PloS one* 14, 3 (2019), e0213668.
[72] Yubao Tang, Ruqing Zhang, Jiafeng Guo, and Maarten De Rijke. 2023. Recent Advances in Generative Information Retrieval. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 294–297.
[73] Richard H Thaler. 2016. Behavioral economics: Past, present, and future. *American Economic Review* 106, 7 (2016), 1577–1600.
[74] Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics* 12 (2024), 1011–1026.
[75] Amos Tversky and Daniel Kahneman. 1985. *The framing of decisions and the psychology of choice*. Springer.
[76] Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, and Fabiana Zollo. 2019. Polarization and fake news: Early warning of potential misinformation targets. *ACM Transactions on the Web (TWEB)* 13, 2 (2019), 1–22.
[77] Ellen M Voorhees. 2019. The evolution of cranfield. *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF* (2019), 45–69.
[78] Ben Wang and Jiqun Liu. 2023. Investigating the role of in-situ user expectations in Web search. *Information Processing & Management* 60, 3 (2023), 103300.
[79] Ben Wang and Jiqun Liu. 2024. Cognitively Biased Users Interacting with Algorithmically Biased Results in Whole-Session Search on Debated Topics. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*. 227–237.
[80] Douglas H Wedell and Jonathan C Pettibone. 1996. Using judgments to understand decoy effects in choice. *Organizational Behavior and Human Decision Processes* 67, 3 (1996), 326–344.
[81] Chunhua Wu and Koray Cosguner. 2020. Profiting from the decoy effect: A case study of an online diamond retailer. *Marketing Science* 39, 5 (2020), 974–995.
[82] Dan Wu, Jing Dong, Li Shi, Chunxiang Liu, and Jiangyun Ding. 2020. Credibility assessment of good abandonment results in mobile search. *Information Processing & Management* 57, 6 (2020), 102350.
[83] Linhai Wu, Pingping Liu, Xiujuan Chen, Wuyang Hu, Xuesen Fan, and Yuhuan Chen. 2020. Decoy effect in food appearance, traceability, and price: Case of consumer preference for pork hindquarters. *Journal of Behavioral and Experimental Economics* 87 (2020), 101553.
[84] Yusuke Yamamoto and Katsumi Tanaka. 2011. Enhancing credibility judgment of web search results. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1235–1244.
[85] Yichao Yuan and Tiaojun Xiao. 2022. Retailer's decoy strategy versus consumers' reference price effect in a retailer-Stackelberg supply chain. *Journal of Retailing and Consumer Services* 68 (2022), 103081.
[86] Reza Zafarani, Xinyi Zhou, Kai Shu, and Huan Liu. 2019. Fake news research: Theories, detection strategies, and open problems. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 3207–3208.
[87] Fan Zhang, Yiqun Liu, Xin Li, Min Zhang, Yinghui Xu, and Shaoping Ma. 2017. Evaluating web search with a bejeweled player model. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 425–434.
[88] Fan Zhang, Jiaxin Mao, Yiqun Liu, Xiaohui Xie, Weizhi Ma, Min Zhang, and Shaoping Ma. 2020. Models versus satisfaction: Towards a better understanding of evaluation metrics. In *Proceedings of the 43rd international acm sigir conference on research and development in information retrieval*. 379–388.
[89] Xichen Zhang and Ali A Ghorbani. 2020. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management* 57, 2 (2020), 102025.
[90] Yan Zhang, Yalin Sun, and Bo Xie. 2015. Quality of health information for consumers on the web: a systematic review of indicators, criteria, tools, and evaluation results. *Journal of the Association for Information Science and Technology* 66, 10 (2015), 2071–2084.
[91] Shanshan Zhen and Rongjun Yu. 2016. The development of the asymmetrically dominated decoy effect in young children. *Scientific reports* 6, 1 (2016), 1–7.
[92] Aljaž Zrnec, Marko Poženel, and Dejan Lavbič. 2022. Users' ability to perceive misinformation: An information quality assessment approach. *Information Processing & Management* 59, 1 (2022), 102739.