Identifying Confirmation Bias in a Search as Learning Task: A Study on The Use of Artificial Intelligence in Education

Marcelo Machado[1], **Jairo F. de Souza**[2], **Sean W. M. Siqueira**[1]

[1]Postgraduate Program in Informatics
Federal University of the State of Rio de Janeiro (UNIRIO)
Rio de Janeiro – RJ – Brazil

[2]Department of Computer Science – Federal University of Juiz de Fora (UFJF)
Juiz de Fora – MG – Brazil

marcelo.machado@edu.unirio.br, jairo.souza@ice.ufjf.br
sean@uniriotec.br

Abstract. *Confirmation bias, the tendency to favor information that supports existing beliefs, can hinder information-seeking, especially in learning contexts where it can perpetuate a one-sided perspective. This paper examines how confirmation bias affects search behaviors among 84 participants learning about AI in education. Participants were divided into Neutral and Biased groups based on their prior attitudes, with the Biased group receiving reinforcing information beforehand. Participants' interactions with the search system were logged, and we analyzed the data for behavioral differences. Results showed that biased participants often completed searches quickly, spending less time engaging with and selecting search results, and issued longer queries. However, other variables showed no statistical difference. Some results contradict other studies on confirmation bias in search, highlighting the complexity of search dynamics in learning contexts and suggesting the need for specialized research into cognitive biases in search as a learning process.*

1. Introduction

Learning is a fundamental aspect of human life, representing a crucial element of developmental psychology (Gagne, 1968) and in the digital age, search tools have become indispensable in facilitating learning. These tools provide instant access to vast amounts of information, enabling learners to explore different domains at their own pace.

However, traditional search systems are generally optimized for ad-hoc tasks (e.g., navigational or fact-finding tasks), they provide less support for searchers working on complex tasks that involve learning (Vakkari, 2016). Learning-centered searches are typically open-ended, aimed at making sense of and understanding the retrieved information. In this scenario, the user's information need is unlikely to be satisfied with a single query. Rather, their interaction with the search system spans longer, encompassing multiple, distinct queries and documents (Otto et al., 2022). This disparity between search systems optimized for single interactions and the multifaceted search scenarios required by learners has given rise to the field of Search as Learning (SAL), studying the nuances of search sessions that are carried out to acquire knowledge and how information systems can evolve to assist these processes (Rieh et al., 2016, Vakkari, 2016).

DOI: 10.5753/sbie.2024.242341

Many factors can influence people's search behavior and performance in search tasks. Over the years, studies have investigated users characteristics (e.g., expertise, topic knowledge, cognitive abilities), system features (e.g., interface, presentation), task attributes (e.g., complexity, goal), learning theories and others (Gimenez et al., 2020, Kelly and Sugimoto, 2013, Machado et al., 2020). Recently, there has been a growing interest in studying the impact of cognitive biases on information-seeking behaviors and their effects on information processing and decision-making (Azzopardi, 2021, Gomroki et al., 2023).

Cognitive biases are systematic deviations in thinking which may lead to errors in judgments and decision-making (Tversky and Kahneman, 1974). As more individuals rely on search systems to access, discover, and consume information for making important life decisions, such as medical, political, social, personal, and financial, investigating and mitigating biases' negative impact on search becomes imperative (Azzopardi, 2021).

Among a wide range of cognitive biases, confirmation bias stands out in the learning context (Machado et al., 2024). Confirmation bias is a tendency to remember, interpret, or search for information in a way that confirms initial beliefs or hypotheses. In other words, it can limit learning to a single perspective.

Although the literature around cognitive biases and confirmation bias has been growing, to our knowledge, they have not been explored in the context of SAL, but only on ad-hoc tasks. On this type of task, searchers know exactly what they are looking for and typically expect a straightforward process. This requires little exploration and generally does not involve the user in meaningful learning compared to the more cognitively demanding searches in SAL contexts (Athukorala et al., 2016).

Research into cognitive biases in search reveals that varying experimental settings can significantly influence outcomes (Azzopardi, 2021). In this study, we examine the effects of confirmation bias within a SAL task. We recruited 84 participants to learn about "The use of Artificial Intelligence (AI) in Education". Our objective is to raise search behavior indicators that may suggest the influence of confirmation bias.

After statistical analysis of users interaction logs, the results show that biased participants often completed searches quickly, spending less time engaging with and selecting search results, and issued longer queries. These are important indicators for creating models capable of identifying and, subsequently, trying to mitigate the negative effects of confirmation bias. However, other variables showed no statistical difference. Some results were contrary to others published on simpler tasks (e.g., the study of Suzuki and Yamamoto (2021)), highlighting the complexity of search dynamics in learning contexts and suggesting the need for specialized research into cognitive biases in SAL.

The remainder of this paper is organized as follows: Section 2 provides a concise background on confirmation bias; Section 3 highlights some important related works; Section 4 details the implied research method; Section 5 presents the experiment results; Section 6 discusses the paper inline with the presented results. Finally, Section 7 concludes the paper and offers insights into future work.

2. Confirmation Bias

A cognitive bias is a systematic pattern of deviations in thinking that occurs in particular situations, leading to perceptual distortion, inaccurate judgment, illogical interpretation, or what is broadly called irrationality (Hirschman, 1983). Cognitive biases might hamper critical thinking and, as a result, the validity of our decisions (Rieh et al., 2016).

The types of cognitive bias are many (Haselton et al., 2015) and they are omnipresent in a wide variety of situations with the potential for adverse effects on the social good. Since the seminal work by Hirschman (1983) over 180 different biases and effects have been identified (Benson, 2016). These have been broadly categorized into four high-level groupings depending on: (i) the amount of information available/presented; (ii) the lack of meaning associated with the information; (iii) the need to act fast; and (iv) what information is remembered or recalled (Benson, 2016).

Users' actions can be negatively influenced by cognitive biases, which can in turn hinder task performance. Detecting these biases can help users become aware of them and make efforts to overcome them when carrying out their tasks (Baeza-Yates, 2018).

Confirmation bias is one of the most discussed biases in the information search process literature (Azzopardi, 2021). It is defined as a tendency to collect information that confirms preexisting beliefs. Emphasizing or seeking evidence that supports these beliefs while dismissing or failing to examine contradictory evidence are habitual behaviors about this bias. Nickerson (1998) points out that the way information is interpreted is also affected by confirmation bias. Individuals seek and interpret evidence according to preexisting beliefs, expectations, and hypotheses. As a result, this bias can lead to poor decisions, as other viewpoints and possibilities may be ignored.

The theory of cognitive dissonance, described by Festinger (1957), states that when an individual holds two or more contradictory beliefs simultaneously or is confronted with new information that contradicts preexisting beliefs, ideas, and values, they experience mental stress or discomfort. This discomfort occurs due to the interference with the consistency between the elements of a cognitive system that, in the absence of conflicts, experiences a state of comfort associated with sensations such as familiarity, ease, and the perception of something good or true. Thus, anything that allows our associative mechanism to function smoothly will also predispose beliefs, which is one of the reasons why confirmation bias manifests.

3. Related Work

Research on cognitive biases in information-seeking and retrieval has grown significantly. Many studies are similar to ours in terms of empirical experiments on cognitive biases in the search process. Due to space limitations, I encourage readers to explore Azzopardi (2021) for a detailed overview. They categorize research by specific biases and search stages, highlighting their impact in areas like health, politics, and web use. While many biases are relevant, our study focuses on confirmation bias.

On confirmation bias, Suzuki and Yamamoto (2021) conducted the study most similar to ours, instructing participants to search for health-related information and comparing the behavior of a manipulated group to a control group. They measured variables such as dwell time on search engine results pages (SERPs), dwell time on web pages, and search session time. However, their task had a straightforward objective, requiring participants to answer a predetermined question, which might limit exploration once the correct answer is found. Overall, the task lacked the learning aspect and associated complexity. Additionally, the researchers did not attempt to understand participants' prior beliefs, relying entirely on the intervention that was supposed to manipulate the participants.

Our work employs a similar methodology but applies a more complex search task on AI in Education. Unlike Suzuki and Yamamoto (2021), we did not explicitly state what participants should do with the acquired knowledge, encouraging greater exploration. Additionally, our approach is quasi-experimental, meaning that the group division was not entirely randomized but primarily based on participants' prior attitudes on the topic, which we assessed through a pre-test. Finally, we also incorporated additional evaluation metrics encompassing variables covering all stages of the search process.

4. Method

The experiment described here was carried out in January 2024[1]. It employs a quantitative research methodology to investigate behavioral differences between two distinct participant groups. This experiment was approved by the research ethics council of the Federal University of the State of Rio de Janeiro.

Participants were divided into two groups based on a pre-test. The test group consisted of those more prone to confirmation bias, identified by their strong pre-existing attitudes on the search topic. Following the method applied in (Kong et al., 2019) and later in (Suzuki and Yamamoto, 2021), we expose these participants to manipulative information that reinforces their beliefs.

4.1. Participants

Participation in our research was limited to people over 18 years old and fluent in Portuguese. Before recruitment, we conducted a power analysis[2] to determine the minimum sample size necessary to detect a statistically significant difference between groups. This analysis used Python, specifically employing the **statsmodels** package with its **TTestIndPower** class. We identified a required sample size of 64 participants, guided by standard parameters widely used in such analyses: Effect Size (d) of 0.8, Alpha Level (α) set at 0.05, and Power ($1 - \beta$) of 0.80, which are commonly accepted in statistical practice.

4.2. Search as Learning Task

We chose "The use of Artificial Intelligence in Education" as the topic for the SAL task. This choice was guided by a thoughtful consideration of the following criteria:

- **Hot-Topic:** The field of Artificial Intelligence (AI) is at the forefront of technological innovation, with advancements such as ChatGPT, driving widespread interest;
- **Diversity of Perspectives:** It encompasses many perspectives, including students, teachers, and stakeholders' viewpoints, privacy, and ethical considerations. This diversity offers a rich terrain for the exploration of varied attitudes;
- **Potential for Engagement without Polarization:** While controversial, it generally avoids eliciting the intense negative responses often triggered by more polarized political discussions;

[1]All data is publicly available at: https://github.com/sal-research-group/xperframe4sal
[2]About power analysis: https://en.wikipedia.org/wiki/Power_of_a_test

- **Interdisciplinary Appeal:** Its broad and interdisciplinary nature is expected to appeal to a diverse participant base, making it pertinent to a wide demographic spectrum; and
- **Educational Significance:** It is an inherently intriguing subject for learning. Its dynamic nature and the ongoing developments in the field present a valuable opportunity for educational engagement.

The task itself was open-ended (or purposeless), meaning the search goal was not explicitly defined; participants were instructed to explore the topic and complete a post-test afterward. This allowed for exploratory search, letting learners stop when they felt sufficiently informed or due to time constraints. According to Soufan et al. (2022, p. 148)' exploratory search conceptual model, these criteria should foster a higher level of exploration. However, given the topic's relevance, many individuals may already have some level of familiarity with it—or, influenced by the Dunning-Kruger effect, they may mistakenly believe they do. Thus, regarding this specific characteristic, we acknowledge that the level of exploration may be somewhat limited. Nevertheless, we maintained this theme to attract a broader pool of volunteers.

4.3. Pre-test

To gauge participants' prior attitudes toward "The use of Artificial Intelligence in Education", we designed a pre-test with questions aimed at indirectly assessing their predispositions. This strategy avoids direct inquiries about their stance, thus reducing the risk of bias and not fully disclosing our study's broader aims. By asking questions like, "How interested are you in the topic of Artificial Intelligence?" and measuring responses on a 5-point Likert scale, we subtly evaluate participants' attitudes.

To assess whether the questions created were capable of capturing the participants prior attitudes, we applied them to 28 volunteers[3] who would not later participate in the official experiment. This assessment included the target question: "Are you for or against the use of Artificial Intelligence in Education?". Subsequently, we conducted a correlation analysis (Pearson correlation coefficient) between this question and others. Questions showing a moderate to high correlation (coefficient greater than 0.5) with the target question were chosen for inclusion in the final questionnaire.

Finally, we selected six questions and incorporated them into an existing set comprising two mandatory queries regarding participants' interest in and knowledge about AI. This process resulted in a comprehensive questionnaire comprising eight questions, each rated on a 5-point Likert scale from -2 to $+2$. Negative scores reflect opposition to using AI in Education, whereas positive scores indicate support. This scale allows for assessing participant attitudes and strengths in their opinions. With a total possible score of ±16 across eight questions, higher scores indicate stronger attitudes. Figure 1 shows the score distribution among the volunteers. Notably, there is a tendency toward a positive attitude regarding the use of AI in Education.

4.4. Group division

Our study is quasi-experimental, meaning the test and control groups will not be randomly divided. The divisions are made based on the learners' scores on the pre-test.

[3]These participants are not included in the set of 84 participants.

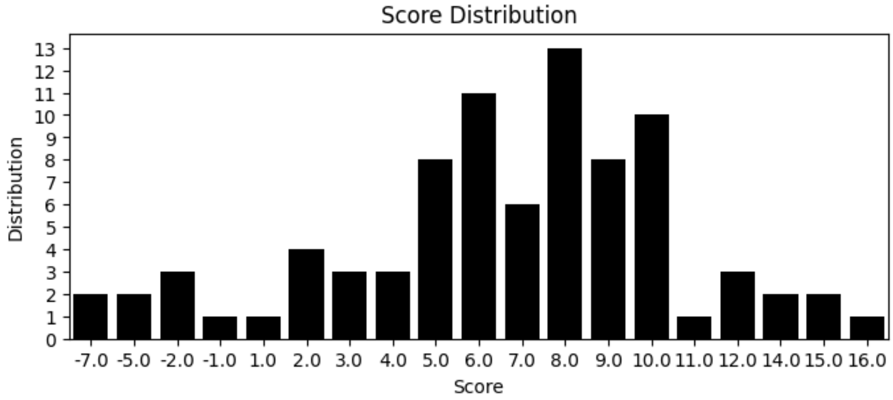

Figure 1. Score distribution of volunteers regarding the use of AI in education.

Given that the recruitment method was the same and the participant profiles were similar to the pre-evaluation volunteers, we decided to take the trend, shown in Figure 1, to divide the groups. Participants with strong attitudes were placed in the test group (*Biased*). Negative scores and scores greater than 9 indicated inclusion in the Biased group. Scores between 0 and 6 placed participants in the *Neutral* group. The remaining scores led to random group assignment. At the end, the Neutral group included 37, while the Biased group included 47. Among the Biased group, 28 had positive attitudes and 19 had negative attitudes about the use of AI in education.

4.5. Procedure

Machado et al. (2024) developed an application-level framework to assist researchers in investigating cognitive biases in search systems. The framework allows instantiating experiments indicating the cognitive bias to be worked on, the search task and the data that must be logged. We used this framework to instantiate our experiment and deployed it on an online server where participants could register and participate. The online user study was performed as follows: (1) Participant registration; (2) Pre-test; (3) Introduction of search topic; (4) Search task; (5) Post-test.

After registering in the system and logging in, learners are presented with a list of experiments available to them. When accessing the experiments page for the first time, learners have to consent to participate in the experiment. On this screen, learners can access the full project document sent to the ethics committee. Learners are redirected to the questionnaire page upon clicking the “accept” button. We included two questionnaires in that stage: the demographic and the pre-test.

After answering the questionnaires, learners are assigned to a task condition according to their assigned group. For the Neutral group, we tried to ensure the instructions were as neutral as possible to avoid a situation where participants became biased due to the instructions. In short, the instructions gave an overview of the growing use of AI in our daily lives and asked learners to search about the relationship between AI and Education, indicating that they would later have to answer an assessment.

In contrast, we offered biased information to the Biased group. For learners with a positive attitude, we presented a video[4] where Alcely Barroso, a renowned teacher on the subject, goes to TEDxMauá to share knowledge about her work and her passion for the duo “education and technology” in solving social problems. The video is in favor of using IA in education. On the other hand, participants with negative attitudes were faced with a video[5] from the BBC News Brazil, warning about the use of AI in a general context.

After this previous intervention, the learner can finally go to the search task. The learner must explore whatever they want on the subject, and when they feel satisfied, they must complete the task using a button inserted in the upper right part of the search interface. Note that no time limit was set for the search task; therefore, learners should decide when they are satisfied with the obtained knowledge.

Learners navigate the way they want by interacting with the search interface, and after deciding to complete the task, they are redirected back to the questionnaire screen. However, at this moment, the questionnaire available is to evaluate what learners have learned. This questionnaire contains some basic questions about the topic, as well as the final question about the stance in favor or against the use of IA in education. The experiment is completed by answering this last questionnaire, and the learner is redirected to the experiments page.

4.6. Measurement data

The variables evaluated are common metrics from experiments on web search behavior (Bateman et al., 2012, Machado et al., 2024, White and Morris, 2007). But another motivation for choosing was to be able to compare our results with those in (Suzuki and Yamamoto, 2021). However, we also included variables about the querying stage, which can also present divergent behavior between users. Then we collected the following data during the search task to analyze the participants’ search behaviors: (i) Number of issued queries; (ii) Query length; (iii) Query stance; (iv) Number of accessed pages; (v) Dwell time on web pages; and (vi) Session time.

Dwell time on search results represents how long each learner spent on search results linked from the SERPs. The search session time is the sum of all sessions for each learner. Regarding the querying stage, we measured the number of queries fired by each learner, the number of characters that make up the string, and the stance related to each query. The number of accessed pages results represents how many search results each learner clicked on during the search tasks.

4.7. Hypothesis

Given that confirmation bias leads individuals to rely on their existing beliefs, we hypothesized that learners with confirmation bias (Biased) would search for information about the use of AI in education less thoroughly than learners without confirmation bias (Neutral). This reduced thoroughness can be measured through interaction variables that indicate participants’ behavior. Therefore, we have tailored our hypothesis as follows:

> Learners with confirmation bias regarding the search topic are expected to exhibit reduced search session time, shorter average dwell times on

[4] At the time of writing, the video was available at https://www.youtube.com/watch?v=et42LUYn18Y
[5] At the time of writing, the video was available at https://www.youtube.com/watch?v=i6xbl9QzIK0

web pages, fewer and shorter queries, and greater alignment between the stances of their queries and their pre-existing attitudes compared to learners without confirmation bias.

4.8. Statistical Analysis

When evaluating our hypotheses, we first assessed the normality of our sample data distribution, which is essential for many statistical tests to produce accurate *p-values* and confidence intervals. If data significantly deviates from normality, results may be misleading, as the statistic tests might not adhere to the expected distribution under the null hypothesis. Subsequently, we checked for homogeneity of variances, as these tests assume that the groups being compared have similar variance. Violating this assumption could lead to errors in estimating Type I and Type II errors, affecting the reliability of the results. Both tests provide a *p-value* to determine normal distribution and variance equality. If the *p-value* falls below the conventional threshold of 0.05, we reject the null hypotheses of normality and equal variance. In such cases, we employed the Mann-Whitney U test[6], a non-parametric test recommended independent samples. To check the normality, we used the Shapiro-Wilk test[7], and for checking the equality of variances, we used Levene's test[8] both from the **scipy** Python library using the **stats** module.

5. Results

This section reports the results of our experiment.

5.1. Participants Characterization

We recruited 156 participants from university email lists and personal social networks without associated costs or benefits. We had to exclude data from 72 participants who did not complete all the experiment steps, leaving us with a total of 84 participants. A high dropout rate was anticipated due to the time-consuming nature of the SAL task, which requires significant dedication and effort, and lack of self-motivation, beyond helping with the experiment, to learn about the topic.

As shown in Figure 2, most participants are aged between 25 and 34 years, followed by those aged 35 to 44. Most are male (non-binary and non-response options were available but not selected). Most have postgraduate education, followed by higher education (no one chose "I prefer not to answer"). Participants reported high or very high internet familiarity; none reported very low familiarity. Regarding occupation, participants were diverse, with the majority being students, teachers and IT-related (e.g., programmers, analysts, system administrators).

5.2. Statistical Analysis

We divided our hypothesis into sub-hypotheses considering each of the variables analyzed. The null hypotheses of normality were rejected in all cases, so we applied non-parametric tests. Table 1 summarizes some of the results and Figure 3 presents it visually.

[6]Mann-Whitney U test: https://en.wikipedia.org/wiki/MannWhitney_U_test
[7]Shapiro–Wilk test: https://en.wikipedia.org/wiki/ShapiroWilk_test
[8]Levene's test: https://en.wikipedia.org/wiki/Levene's_test

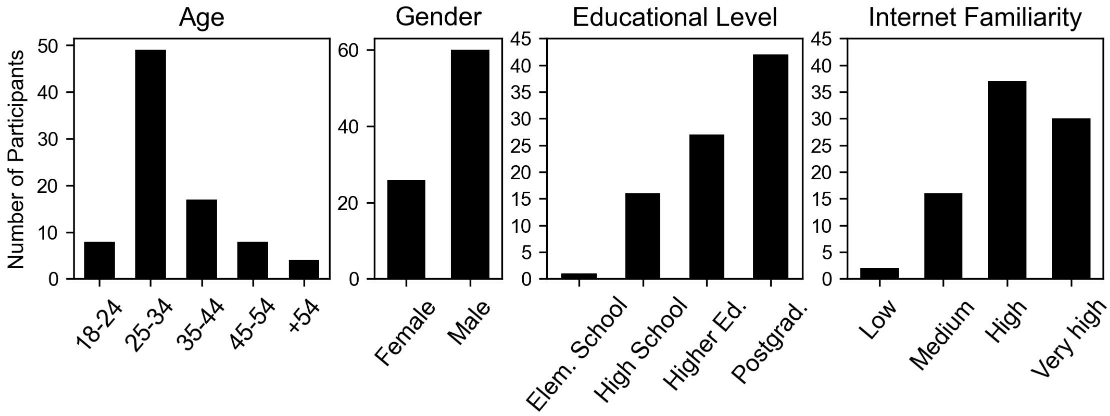

Figure 2. Participant demographic characteristics.

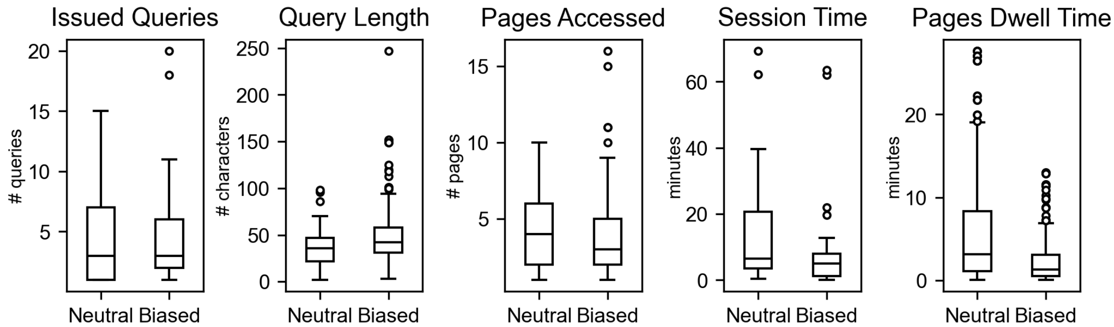

Figure 3. Numerical difference between the two groups considering each of the evaluated metrics.

The first boxplot in Figure 3 compares the number of queries issued by the two groups. The median number of queries is quite similar between the groups. However, the IQR (the middle 50% of data) is broader in the Neutral group, indicating a wider variation in the number of queries. The Biased group has a slightly narrower IQR, indicating more consistent query numbers among its members. Table 1 shows that the *p-value* for this metric exceeded 0.05 in the Mann-Whitney U test, failing to reject the null hypothesis of no significant difference between the groups. Therefore, the data did not provide sufficient evidence to support a difference between the groups.

The second boxplot compares query length, showing a significantly higher median for the Biased group compared to the Neutral group, suggesting they issued longer queries on average. Table 1 shows a *p-value* less than 0.05, rejecting the null hypothesis. This indicates a statistically significant difference. However, the difference was contrary to what we hypothesized, as we expected that the Neutral group would issue longer queries.

The third boxplot compares the number of pages accessed, with the Neutral group showing a slightly higher median and greater variability, suggesting that, on average, the Neutral group accessed more web pages. Table 1 shows a *p-value* above 0.05, meaning the null hypothesis could not be rejected.

The fourth boxplot compares session time. The Neutral group has a higher median session time than the Biased group, with a significantly wider range of session times. Table 1 shows a *p-value* less than 0.05, allowing us to reject the null hypothesis of no

Table 1. Normality and equal variances tests of the analyzed variables.

Issued Queries			
Groups	Shapiro-Wilk's test	Levene's test	Mann-Whitney U test
Biased	0.786, p=0.000, Not normal	0.013, p=0.911, Equal variance	826.0, p-value: 0.695, rejected
Neutral	0.817, p=0.000, Not normal		
Query Length			
Groups	Shapiro-Wilk's test	Levene's test	Mann-Whitney U test
Biased	0.845, p=0.000, Not normal	11.770, p=0.001, Not equal variance	14227.0, **p-value: 0.0002, accepted**
Neutral	0.967, p=0.001, Not normal		
Pages Accessed			
Groups	Shapiro-Wilk's test	Levene's test	Mann-Whitney U test
Biased	0.834, p=0.000, Not normal	0.114, p=0.737, Equal variance	768.5, p-value: 0.734, rejected
Neutral	0.895, p=0.003, Not normal		
Session Time			
Groups	Shapiro-Wilk's test	Levene's test	Mann-Whitney U test
Biased	0.532, p=0.000, Not normal	2.128, p=0.149, Equal variance	827.0, **p-value: 0.017, accepted**
Neutral	0.763, p=0.000, Not normal		
Web Page Dwell Time			
Groups	Shapiro-Wilk's test	Levene's test	Mann-Whitney U test
Biased	0.786, p=0.000, Not normal	36.709, p=0.000, Not equal variance	21206.0, **p-value: 7.608e-10, accepted**
Neutral	0.737, p=0.000, Not normal		
Query Stance			
Groups	Shapiro-Wilk's test	Levene's test	Mann-Whitney U test
Biased	0.416, p=0.000, Not normal	7.579, p=0.006, Not equal variance	15179.5, p-value: 0.440, rejected
Neutral	0.599, p=0.000, Not normal		

significant difference in session time between the groups, supporting our hypothesis.

The last boxplot compares web page dwell time. The Neutral group has a higher median dwell time and a significantly wider range of dwell times compared to the Biased group. Table 1 shows a *p-value* less than 0.05, allowing us to reject the null hypothesis of no significant difference between the groups, supporting our hypothesis.

Finally, we used the ChatGPT[9] to evaluate the query stance variable. We used the `gpt-4` model with a `temperature` setting of 0.2 and `top_p` set to 1.0 to minimize randomness and ensure consistency in the results. ChatGPT was chosen for this task because of its ability to process natural language inputs with high contextual understanding, making it particularly suited for identifying nuanced stances in diverse queries. We crafted the following prompt, where the placeholder `{{query}}` was replaced for each query submitted by learners in the system:

```
Given the following examples:

Sentence: Benefits of using AI in Education
Stance: Positive
Sentence: Harms of using AI in Education
Stance: Negative
Sentence: Use of AI in Education
Stance: Neutral

Complete the stance of the following sentence:
Sentence: {{query}}
Stance:
```

[9]OpenAI API: https://openai.com/index/openai-api

Table 2 presents the quantitative information about this analysis. Of the 228 different queries, 195[10] was classified as neutral. Additionally, the model included, on its own, the `Irrelevant` label.

Table 2. Number of queries triggered according to each stance.

Stance	Example	Number of queries	Neutral	Biased
Neutral	Artificial Intelligence in Education	195	85	118
Negative	the dangers of AI in education	13	4	9
Positive	advantages of artificial intelligence	12	5	7
Irrelevant	cross-multiplication	8	7	1

As shown in Table 1, the *p-value* exceeded 0.05 in the Mann-Whitney U test, failing to reject the null hypothesis of no significant difference between the groups according to the query stance variable.

6. Discussion

In the previous section, we divided our hypothesis according to each analyzed variable. Although absolute differences were observed between groups for all variables, only time spent on web pages, total session time, and query length showed statistically significant differences. On the time variables, these results contrast with those reported in (Suzuki and Yamamoto, 2021), where no statistical differences were found. This highlights the complex dynamics of this domain and the challenges in generalizing results and developing models to predict and mitigate cognitive biases. One possible factor contributing to this contradiction is the complexity of the task. In their study, users accessed health information to answer pre-known questions, limiting their actions due to the closed-ended (or purposeful) nature of the task. A purposeless task, as ours, should allows for broader exploration. By indicating the expected outcome, it could limit interactions once learners find a single answer that aids in addressing the final task. On the other hand, the open-ended nature of our task likely led to the formulation of queries with more neutral stances. If the task had been explicitly defined, such as justifying a stance, it might have resulted in more polarized queries. Future research should test this hypothesis.

We found statistical difference between the groups on the query length variable, but in the opposite direction of what we expected. Upon reviewing the logs, we observed that longer queries were often natural language questions, resembling those used in chat interfaces. Additionally, some participants entered website links directly into the search bar. We suspect that factors like search skills, prior knowledge, and the shift from traditional search tools to chat-based interfaces may have influenced these results.

We observed a limited variety of queries, as out of 427 issued, only 228 were unique in a simple string comparison (a semantic comparison would show even less diversity). The query "use of artificial intelligence in education" was by far the most issued; that is, the learners went straight to the point of what the search task was referring to. In more realistic SAL situations, users would not have this "anchor" (about the anchoring bias), and we could notice the influence of ASK[11]. In other words, the task condition we

[10] This number represents unique queries across all groups, which justifies the number of neutral queries being 195, while the sum of the number of neutral queries across groups is 203.

[11] ASK stands for Anomalous State of Knowledge, where users struggle to articulate their information needs due to a lack of familiarity with the subject matter (Belkin, 1980).

presented is still shallow for fully understanding the dynamics between biases and a more complex SAL task. However, this is justified by the complexity of creating these more realistic scenarios in observable environments.

It is worth noting that doing research on bias is inherently challenging. Awareness of the effects of these biases is exactly the first step toward mitigation. Therefore, experiments cannot be explicit about this motivation as it can completely interfere with behaviors and results. Cognitive biases will likely decrease or disappear if a task or context stimulates more analytic information processing, for example, by triggering high personal accountability or critical thinking in the learner (Rieger et al., 2021). In any case, external factors could modify the context in which learners operate, potentially influencing their natural behavior. For instance, the main motivation of the participants of our experiment was generally "to help", that is, there is not a strong connection regarding the need to search. Thus, interactions in general, regardless of the task condition, tend to be shallower. This observation is supported by our analyzed time variables. While extensive research might take days, our experiment shows that most sessions conclude within minutes despite we do not set any time restrictions.

In response to our research question, the results indicate that original expectations about the impact of confirmation bias on the SAL task have not been fully confirmed since we could not observe statistical differences in all the evaluated metrics.

7. Concluding Remarks

In this paper, we have described an experiment to identify search behaviors in consideration of confirmation bias. To divide users into groups with and without confirmation bias, we provided the participants with prior information to manipulate their impressions about the SAL topic. We then analyzed the logs of their search and browsing.

Our results demonstrated that confirmation bias can influence learners' search behavior and highlight that the complexity of the task impacts this behavioral difference. However, not all variables demonstrated statistical difference. We suggest that these results raise important indicators so in future endeavors these could be used to create models to mitigate the negative effects of this confirmation bias.

In future work, we plan to expand the list of variables analyzed to include the complexity and type of queries, the use of search tool feedback such as "people also ask", and a variable indicating whether learners are accessing or not only search results that align with their pre-existing attitude on a topic. We also aim to evaluate the impact of other cognitive biases, such as anchoring and availability biases, and explore compound effects where one bias influences another. Additionally, with the increasing use of chat-based search interfaces in the digital age, we intend to study their impact on search behaviors and learning outcomes. Finally, we will conduct a more thorough analysis of how these biases affect learning outcomes, aiming to develop strategies to mitigate their negative effects and enhance the overall learning experience.

We hope that this work serves as inspiration for investigating the many complexities of cognitive biases in SAL, encouraging further research to uncover deeper insights and develop effective strategies to enhance search behaviors and learning outcomes. By understanding and addressing these biases, we can create more equitable and efficient learning environments that better support the diverse needs of learners.

Acknowledgment

This study was partially funded by the 'National Council for Scientific and Technological Development – CNPq' (Process 305436/2021-0) and by the Coordination of Superior Level Staff Improvement (CAPES)—Support Program for the Dissemination of Scientific and Technological Information (PADICT) and CAPES Periodical Portal)—Finance Code 001.

References

Athukorala, K., Głowacka, D., Jacucci, G., Oulasvirta, A., and Vreeken, J. (2016). Is exploratory search different? a comparison of information search behavior for exploratory and lookup tasks. *Journal of the Association for Information Science and Technology*, 67(11):2635–2651.

Azzopardi, L. (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*, CHIIR '21, page 27–37, New York, NY, USA. Association for Computing Machinery.

Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM*, 61(6):54–61.

Bateman, S., Teevan, J., and White, R. W. (2012). The search dashboard: how reflection and comparison impact search behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1785–1794.

Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian journal of information science*, 5(1):133–143.

Benson, B. (2016). Cognitive bias cheat sheet. `https://betterhumans.pub/cognitive-bias-cheat-sheet-55a472476b18`. [Online; accessed July-2023].

Festinger, L. (1957). *A theory of cognitive dissonance*, volume 2. Stanford university press.

Gagne, R. M. (1968). Contributions of learning to human development. *Psychological review*, 75(3):177.

Gimenez, P., Machado, M., Pinelli, C., and Siqueira, S. (2020). Investigating the learning perspective of searching as learning, a review of the state of the art. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 302–311, Porto Alegre, RS, Brasil. SBC.

Gomroki, G., Behzadi, H., Fattahi, R., and Salehi Fadardi, J. (2023). Identifying effective cognitive biases in information retrieval. *Journal of Information Science*, 49(2):348–358.

Haselton, M. G., Nettle, D., and Murray, D. R. (2015). The evolution of cognitive bias. In *The Handbook of Evolutionary Psychology*, pages 1–20, Online. John Wiley & Sons, Ltd.

Hirschman, E. C. (1983). Judgement under uncertainty: Heuristics and biases. *JMR, Journal of Marketing Research (pre-1986)*, 20(000002):217.

Kelly, D. and Sugimoto, C. R. (2013). A systematic review of interactive information retrieval evaluation studies, 1967–2006. *Journal of the American Society for Information Science and Technology*, 64(4):745–770.

Kong, H.-K., Liu, Z., and Karahalios, K. (2019). Trust and recall of information across varying degrees of title-visualization misalignment. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13.

Machado, M., Assis, E. C., Souza, J. F., and Siqueira, S. W. M. (2024). A framework to support experimentation in the context of cognitive biases in search as a learning process. In *Proceedings of the 20th Brazilian Symposium on Information Systems*, SBSI '24, New York, NY, USA. Association for Computing Machinery.

Machado, M., Gimenez, P., and Siqueira, S. (2020). Raising the dimensions and variables for searching as a learning process: A systematic mapping of the literature. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 1393–1402, Porto Alegre, RS, Brasil. SBC.

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

Otto, C., Rokicki, M., Pardi, G., Gritz, W., Hienert, D., Yu, R., von Hoyer, J., Hoppe, A., Dietze, S., Holtz, P., et al. (2022). Sal-lightning dataset: Search and eye gaze behavior, resource interactions and knowledge gain during web search. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, pages 347–352.

Rieger, A., Draws, T., Theune, M., and Tintarev, N. (2021). This item might reinforce your opinion: Obfuscation and labeling of search results to mitigate confirmation bias. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, HT '21, page 189–199, New York, NY, USA. Association for Computing Machinery.

Rieh, S. Y., Collins-Thompson, K., Hansen, P., and Lee, H.-J. (2016). Towards searching as a learning process: A review of current perspectives and future directions. *Journal of Information Science*, 42(1):19–34.

Soufan, A., Ruthven, I., and Azzopardi, L. (2022). Searching the literature: an analysis of an exploratory search task. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 146–157.

Suzuki, M. and Yamamoto, Y. (2021). Analysis of relationship between confirmation bias and web search behavior. In *Proceedings of the 22nd International Conference on Information Integration and Web-Based Applications & Services*, iiWAS '20, page 184–191, New York, NY, USA. Association for Computing Machinery.

Tversky, A. and Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131.

Vakkari, P. (2016). Searching as learning: A systematization based on literature. *Journal of Information Science*, 42(1):7–18.

White, R. W. and Morris, D. (2007). Investigating the querying and browsing behavior of advanced search engine users. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 255–262.