Belief Dynamics in Web Search

Ryen W. White
Microsoft Research
One Microsoft Way
Redmond, WA 98052 USA
ryenw@microsoft.com

Keywords:
Belief dynamics, Web search, Cognitive biases, Information retrieval

Word count: 11536

Abstract

People frequently answer consequential questions, such as those with a medical focus, using Internet search engines. Their primary goal in performing these searches is to revise or establish beliefs associated with one or more outcomes. Search engines are not designed to furnish answers, and instead provide a ranked list of documents, some of which contain answers. People's prior beliefs about answer likelihoods can affect result selection decisions, and when these selections are aggregated across searchers they can skew machine-learned result rankings. Research on information retrieval has targeted key aspects of information access such as query formulation, result relevance, and search success. However, there are important unanswered questions surrounding how beliefs—and potential biases in those beliefs—affect search behaviors and how beliefs are shaped by search engine use. In this article, we report on a study examining changes, or dynamics, in beliefs during Web search. To understand belief dynamics, we focus on a balanced set of yes-no medical questions (e.g., "is congestive heart failure a heart attack?"), with consensus answers from physicians. We show that: (1) pre-search beliefs are affected only slightly by searching and any changes in belief are more likely to move toward positive (*yes*); (2) pre-search beliefs affect search behavior; (3) search engines can shift some beliefs by manipulating the relative ordering and availability of answers in the results, but strongly-held beliefs are both difficult to move using uncongenial information and can be counterproductive, and; (4) search engines may exhibit near-random answer accuracy, but they may be able to help searchers attain higher accuracy if engines consider factual correctness. Among other things, our findings suggest that search engines must attain, represent, and utilize correct answers in result ranking, and develop methods to encourage searchers to shift strongly-held but factually-incorrect beliefs.

1. Introduction and Background

Internet search engines are a primary mechanism by which laypeople seek answers to questions, reduce uncertainty, and learn about the world. Irrespective of whether search engines are currently designed to provide direct answers, the significant fraction of search queries that are direct questions[1] suggests that searchers expect this functionality and that they may associate result ordering with a ranking of outcomes by likelihood of occurrence (White and Horvitz, 2009). The objective in question-answering is to shape *beliefs* (i.e., a psychological state in which an individual holds a proposition or premise to be true) and resolve uncertainty. The means by which this occurs has been the focus of some previous study in the information science community. Brookes (1980) demonstrated that information modifies knowledge structures, which play a central role in determining beliefs (Kimble, 2013). The cognitive processes behind search activity have been studied in detail, but studies have focused on the formulation of information

[1] Recent research suggests that around 6% of queries issued to search engines are formulated as directed questions (Rose and Levinson, 2004) and 2% of search engine queries are yes-no questions (White, 2013).

needs, changes in those needs during the search process, and their impact on search behavior (Belkin et al., 1982; Ingwersen, 1994; Marchionini, 1995; Taylor, 1968). Although uncertainty has been investigated in studies of information seeking (Bates, 1989; Dervin, 1983; Kuhlthau, 1991; Kuhlthau, 1993; Wilson et al., 2002), the uncertainty studied typically surrounds a user's state of knowledge, which may be high at the outset but reduces over time, as they review information (although not always (Wilson et al., 2002)).

In probability theory and statistics, Bayes' theorem can be applied to revise a probability estimate for a hypothesis in light of additional evidence (Bayes, 1763). The theorem expresses how a subjective degree of belief should rationally change to account for that evidence. However, humans can be irrational agents (Elster, 1979). Research on bounded rationality (Ariely, 2008; Simon, 1991) and hypothesis testing strategies (Baron, 2007; Wason, 1960) suggests that people may not act rationally (e.g., in pursuing factually-incorrect information) as they answer questions or solve problems. Biases in people's beliefs can draw them toward irrational outcomes, significantly affecting their judgment and decision making. These effects have been studied for decades in psychology (Gigerenzer and Todd, 2000; Klayman and Ha, 1987; Tversky and Kahneman, 1974). Belief dynamics has also been researched extensively (Anderson, 1981; Hogarth and Einhorn, 1992; Tversky and Kahneman, 1974), including how the dynamism is affected by factors such as information type (e.g., textual vs. numeric) and the presentation order of experimental stimuli. Research on cognitive dissonance and selective exposure to attitude-supporting information (Eagly and Chaiken, 1993; Fischer et al., 2011; Frey, 1986; Hart et al., 2009) also suggests that information seekers favor supporting information driven by both accuracy and defense (confirmatory) motivations. Despite the close relationship between beliefs, biases, and information seeking, little research has been performed to understand the role of these factors in common retrieval settings such as Web search. Among the findings of the detailed investigation presented in this article, we show that search engines offer little assistance in helping searchers form factually-correct beliefs, and that searcher beliefs can often be shifted using manipulations in the availability and rank ordering of answer pages in search results, but not for strongly-held beliefs, and this bias can be counterproductive.

Information retrieval researchers and search engine designers have considered biases but in ways that are largely unrelated to cognition. That research has focused on topically-skewed result sets (Fortunato et al., 2006; Goldman, 2006), or the impact of rank position, captions, or domain preferences on search-result selection (Ieong et al., 2012; Joachims et al., 2007; Yue et al., 2010). Since a changed belief state may be one desired outcome of searching (another being a reinforcement of existing beliefs), we need to better understand the mechanisms by which beliefs evolve, the role of search engines in affecting these changes, and the impact of beliefs on search behavior. The latter is particularly important in search-result ranking applications, given that behavioral signals are often mined in the aggregate to estimate result relevance for

future queries (Agichtein et al., 2006; Joachims, 2002). Ranking algorithms may learn skewed result lists (biased toward a particular perspective, misaligned with reality) if many users share a particular belief and behave irrationally by preferring information aligned with erroneous outcomes such as myths or common misconceptions (Cho and Roy, 2004; Goldman, 2006). For example, a parent using a search engine to seek an objective answer to the question "can vaccines cause autism?" may encounter a biased result list learned from aggregated user behavior even though there is no scientific evidence of a link (Rochman, 2011). For controversial topics such as this and others, recent studies have found that people may benefit from exposure to diverse opinions (Mankoff et al., 2011; Munson and Resnick, 2010). Our focus here is on a set of non-controversial, yes-no questions in the medical domain with a known correct answer, enabling us to carefully examine belief dynamics during the search process and answer accuracy afterwards.

The connection between beliefs, biases, and information retrieval was noted in a recent study of yes-no questions in the medical domain (White, 2013). That study focused on recalled beliefs in retrospective surveys and aggregated search behavior mined from search logs, with no opportunity to elicit information from searchers about their beliefs immediately preceding and succeeding the search. In that study we showed in a naturalistic setting that searchers favored *yes*-related search content significantly more than *no*-related content, and appeared to frequently settle on factually incorrect answers (around 50% of the time). However, the retrospective nature of the log-based analysis and the fact that we were unable to identify the searchers (per the terms of use under which the search logs were collected), meant that we could not ask searchers directly about their beliefs and how they were revised by searching. In that study, we were also unable to control the availability and rank ordering of answer content in the results with a view to affecting post-search beliefs and understanding if and when these beliefs where malleable.

In this article we report on a user study that complements the log-based study where we collect belief estimates from participants *at search time*. We elicit answers to a balanced set of 36 health-related yes-no questions, a subset of questions mined from the logs of Microsoft Bing search engine as part of our earlier study (White, 2013). These questions were then assigned to 344 remote participants, and their answers to those questions were captured immediately before they reviewed the list of the top 10 search results from the engine and immediately after reviewing the result list. We focused on the medical domain given its importance and prior evidence of a link between the presentation order of search results and searchers' self-reported beliefs about outcome likelihoods (White and Horvitz, 2009), as well as concerning connections between the promotion of serious conditions and negative emotional outcomes (Lauckner and Hsieh, 2013). We obtain caption and page answer judgments for the search results from third-party judges and consensus answers from practicing physicians. We assume that the consensus answer is correct, and that because there is expert agreement that the questions are non-controversial.

The study answers the following research questions:

1. How do beliefs change as a result of search?
2. What effect do pre-search beliefs have on search behavior and search outcomes? and
3. To what extent can beliefs be shifted using variations in answer availability and rank ordering if search engines knew the correct answer, and how does this vary with belief strength?

Based on the answers to these questions, we argue that search engines need to develop methods to derive and promote accurate answers for direct question queries, either through ranking or displaying answers directly, as well as considering searchers' prior beliefs about outcomes (and the strength of those behaviors) and their desire to validate them. Although search engines may consider topical interests and resource preferences for applications such as personalized search (Pitkow et al., 2002; Teevan et al., 2005), the nature of searcher *beliefs* in particular outcomes and biases that may both skew those beliefs and affect their willingness to revise them can have a marked impact on search activity and answer accuracy.

The remainder of this article is structured as follows. In Section 2 we describe the user study that we performed to gather answers to our set of yes-no questions from remote participants, and manipulations of answer availability and answer ranking to assess the extent to which beliefs could be altered by search engines (perhaps to direct users with strongly-held, but factually-incorrect beliefs to the correct answer). Section 3 describes the findings of our study. In Section 4 we discuss our findings and their implications, and we conclude by summarizing our contributions and opportunities for future work in Section 5.

2. The Study

In this section, we describe the user study that we performed to answer each of our three research questions. We first summarize the experimental methodology and then we describe some core components of the study including the data (questions, consensus answers, and answer labels for individual captions and search results), the judgment interface, and the controlled manipulation of answer availability and relative ordering that was performed as part of the experimental design.

2.1 Methodology

Participants took part in the study remotely via a Web browser. They could select the human intelligence task from a dashboard provided by Clickworker.com, which offers crowdsourced services under contract to Microsoft and other companies. Participants resided in the United States and were required to be fluent in English. A preview task was offered so that participants better understood the nature of the task before starting; this also served as a practice task for those who decided to proceed. There were a maximum of 36 tasks per searcher, presented in a randomized order. We required answers to each task from 20 participants. Participants were asked to provide their pre-search belief about the answer to the question (on a five-point scale: *yes*, *lean yes, equal* (unsure), *lean no,* and *no*), review a ranked list of search results from the

Microsoft Bing search engine, and then provide their post-search belief about the answer to the question (on the same five-point scale). A participant could abandon a yes-no question at any of the three stages and they were not required to complete all 36 questions. In addition to performing the study with the original (unmodified) top 10 results from Bing, we also ran a set of additional experiments where we purposely modified the relative ordering and availability of answers in the top 10 results to (a) isolate the effect of pre-search beliefs on search behavior and outcomes, and (b) understand if and how beliefs could be changed via result manipulations. Across all experiments, 344 participants provided at least one answer, with on average 18.8 answers per participant (median=18). Since the study was conducted remotely and available to many participants simultaneously, we could not require that people answered all 36 questions available to them. Participants could quit before the set of 36 tasks are completed, or other participants could grab the remaining tasks before they had a chance to complete the set (i.e., a participant could not lock the full set of tasks). Of the 344 participants, 87 (25.2%) answered all questions.

2.2 Data: Questions, Answers, and Content Labels

We now describe the questions that we selected for our study and the process by which we obtained answers to them from trained medical professionals. We focus on yes-no questions in this study since their answers vary along a single dimension (from *yes* to *no*), provide both clear insight into searchers' beliefs and belief dynamics as a result of search, as well as allowing answers to be clearly defined given expert opinion. Although these questions were derived from search logs, they could also be sourced from other venues such as retrospective surveys or crowdsourcing.

2.2.1 Questions

We automatically extracted health-related yes-no questions from a random sample of search logs of queries issued by users of the Microsoft Bing search engine during September 2012. Yes-no questions are an interrogative construction where an answer of *yes* or *no* is required. We automatically analyzed the search logs to extract such questions asked as queries using variants of "be," "have," "do", or a modal verb (e.g., "can," "will"). We also created a list of stop phrases (e.g., "do not call," "do it yourself," "will smith") to remove frequent false positives from our data. Yes-no questions comprised 2% of the query volume on Bing, demonstrating both that they are an important class of queries and that people do indeed seek direct answers to their questions from search engines (an additional important motivation for this research). To be included in our analysis, questions had to be issued by at least 10 users over the two-week period and had to have the same top-10 results and the same *captions* (titles, snippets, URLs) across all instances of the query, to remove these factors as a potential source of variance in the study. The degree of medical intent in each question was then labeled using a using a proprietary classifier provided by Bing. Those that had strong medical intent were selected for inclusion in the study. The questions chosen include "Do food

allergies make you tired?", "Is congestive heart failure a heart attack?", and "Can aspirin cause blood in urine?" More details on the question extraction and medical intent labeling methodologies is provided in our previous work (White, 2013). A set of 1000 yes-no questions that met the criteria outlined in this section was randomly selected as an initial set in our analysis. This set was later substantially reduced in size to make the answering task more manageable for our experimental participants.

2.2.2 Answers

An important part of our investigation was to attain correct answers to our yes-no questions. The correct answers were central to our measurement of answer accuracy before and after reviewing the search results provided (as is covered later in Section 3.4). Two practicing physicians reviewed the questions and each provided an answer for each question. The physicians were recruited through an external vendor and compensated financially for their labor. They worked independently and there was no opportunity for discussion to resolve disagreements. Physicians were asked to provide an answer on the following four-point scale: *yes*, *50/50*, *no*, and *don't know*. In answering the questions, they were asked to think of the most common scenario or circumstances that could apply when a searcher types such a question on the Internet. The 50/50 rating was only to be used if: (i) there really is an equal split between *yes* and *no* in the most common scenario or circumstances, and/or (ii) more information would certainly be needed to provide a definitive answer.

The percent of overall agreement between the physicians across all four answer options for the 1000 questions was 72.2%. The Cohen's free-marginal kappa (κ) inter-rater agreement, considering chance agreement between raters is 0.630, signifying substantial agreement. If we only consider the questions where both judges provided a *yes* or *no* response, the percent of overall agreement rises to 83.4%, with κ=0.668. There were 674 questions for which the two judges agreed on *yes* or *no* as the answer.

We use a randomly-selected subset of these 674 questions with consensus agreement to maintain a manageable answering workload for our user study participants. We restricted the question set to 36 questions: 18 questions where the correct (physician consensus) answer was *yes* and 18 questions where the correct (physician consensus) answer was *no*. Pilot testing showed that this constituted 30 minutes of answering effort, which we deemed to be a reasonable limit given the intensity and repetitious nature of the search task, without tiring our participants and potentially harming the quality of the answers they provided.

2.2.3 Content Labels

So that we could better understand the nature of the information that our participants engaged with and manipulate the result rankings accordingly (e.g., to place *yes*-oriented content higher in the ranking than *no*-oriented content), we also recruited a separate pool of crowdsourced judges to label answer presence in the top-ranked captions and each of the search results.

Judges came from a pool provided under contract to Microsoft by Clickworker.com. The task required judges to assess whether a caption presented on a result page suggested an answer to the current yes-no question. Judges were provided with a yes-no query such as "*Can flonase make you tired?*" and a single caption. The caption may answer the question with yes or no directly (e.g., "*... can flonase make you tired? Yes ...*") or suggests an answer indirectly (e.g., "*... flonase is unrelated to tiredness ...*"). Captions can also contain contradictory answers and no answer. Judges were required to review the caption with respect to the question and rate whether it contained *yes*-only, *no*-only, both *yes* and *no*, or *neither* content. Figure 1 shows examples of each of the caption labels for a variety of queries. These examples are taken from the guidelines provided to judges to help them understand the judgment task and the nature of the judgments that we expected from them.

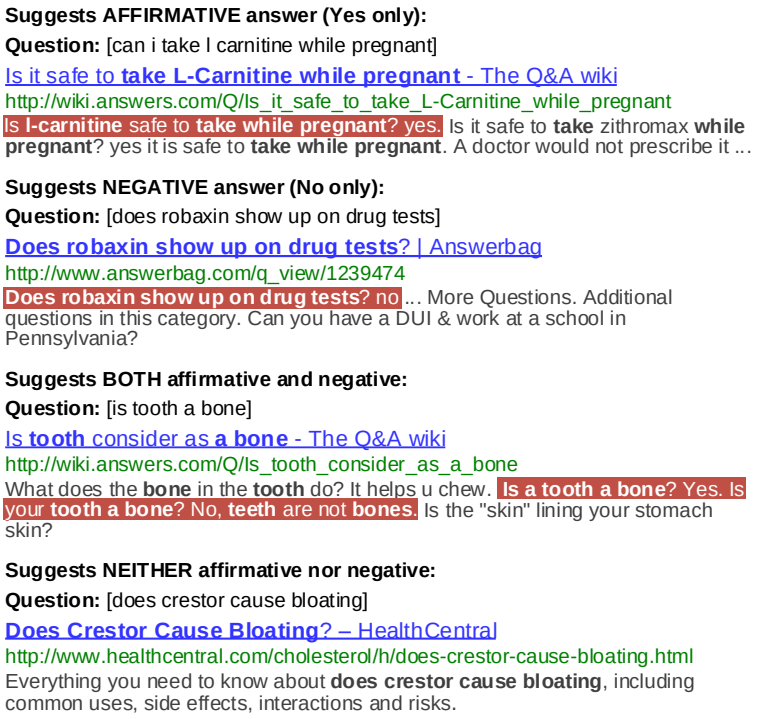

Figure 1. Examples of each of the caption ratings. The answer text is highlighted in the first three captions.

Between three and five judges rated each of the 360 captions under consideration (i.e., 36 yes-no questions × 10 captions per question, one caption for each of the top 10 results) to achieve a consensus comprising at least three judges with the same rating. In total, consensus was achieved for 97% of captions, with 88% of captions attaining agreement with only three judges. In our analysis, we only used those captions with agreement between three judges.

We employed a similar methodology to judge the full text of each of the results returned by the search engine. The content of the judged pages was extracted from the Bing search engine index around the same time as the query logs were collected. This allowed us to present content for each page to our judges, preventing content dynamics over time from potentially biasing judgments. Judges were sourced from the same pool as the caption judging, with at least three judges, and up to five to reach consensus, as with the captions. We found that there was mostly strong agreement between the captions and the landing pages (88%), with the following exceptions: (1) caption label was *neither*, page label was *yes*, *no*, or *both* (7%), and (2) caption label was *yes-only* or *no-only* and page contained both *yes* and *no* (5%).

2.3 Judgment Process

To gather answers to the yes-no questions from user study participants, we created a judgment interface in HTML. The interface was accessed through a Web browser and comprised three phases: (1) pre-search belief capture, (2) result list review, and (3) post-search belief capture. Only one phase was visible on the interface at any time and participants had to click a button to progress through the three phases. The task required that participants answer a yes-no question using a list of search results. Participants were instructed that the goal in reviewing the results was to answer the yes-no question, mimicking a real search scenario. Answers were captured on a five-point scale with response options: *yes*, *lean yes*, *equal* (unsure), *lean no*, and *no*.

Figure 1 presents screenshots of the three phases of the judgment interface that participants used to provide their pre-search beliefs for the yes-no questions, review the search results (including being able to click on the results and review the landing-page content, as needed), and provide the post-search beliefs after they reviewed the results.

For each of the yes-no questions they answered, participants did the following:

- *Phase 1: Rated their prior belief about the outcome (before reviewing the results).* Participants were presented with the yes-no question and asked to indicate on a five-point scale their *prior* belief about the outcome, based on their own knowledge. Participants selected a radio button corresponding to their belief and could only select one option.
- *Phase 2: Reviewed a ranked list of search results, clicking on results as needed.* Participants were presented with a search result list and asked to review the list, and select results as required. The list was described only as a "ranked list of search results" and no indication was given about how the results were generated. This is particularly important for the manipulations described later in the article, where we did not want to draw undue attention to any result changes.

- *Phase 3: Rated their posterior belief about the outcome (after reviewing the results).* Participants were asked to provide their belief in the outcome of the search after reviewing the results presented to them.

Participants were provided with task guidelines. We did not employ a qualification task for these experiments as is common practice in crowdsourcing studies; since we were focused on participants' beliefs there was no gold standard against which to compare their beliefs. We used judge speed tracking to identify outlier judgments which may be indicative of low quality (e.g., erroneous judgments might happen rapidly). Judgment times appeared reasonable (median time=25.81s, average=41.36s) and we did not have to remove any participants from the study for erroneous answering practices.

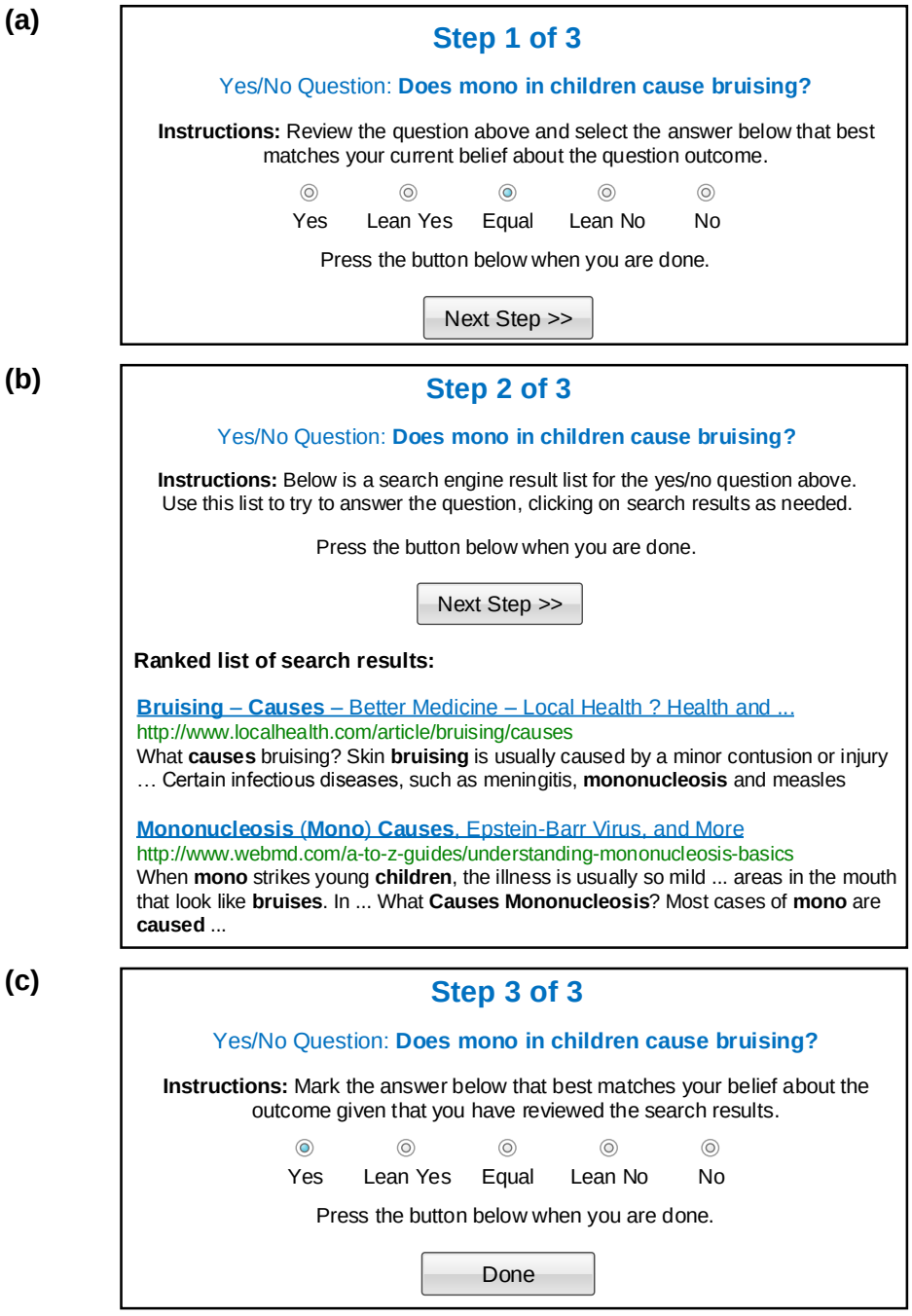

Figure 2. Screenshots of the judgment interface: (a) capturing the pre-search answers, (b) allowing for the review of search results, and (c) capturing the post-search answers.

2.4 Manipulating Answer Availability and Relative Rank Ordering

As stated earlier, in addition to studying the impact of the results as generated by the search engine, we also explored the effect of a number of variations in the availability and rank ordering of the answer pages in the result list. Since our methodology affords the collection of belief ratings from participants at search

time, we were also able to experiment with a variety of result manipulations to influence searchers' post-search beliefs via basic mechanisms that the search engine controls (i.e., the ranked list of results). Being able to affect predictable changes in post-search beliefs could be an important part of directing searchers to the correct answers in next-generation search engines armed with an amplified awareness about factual correctness in the world in which they operate. We wanted to understand the extent to which these changes were attainable via ranking adjustments alone, without the need for special interface treatments, and the extent to which searcher's prior beliefs affect their receptiveness to different result manipulation strategies.

We varied the yes-no outcomes for each of the experiments with three variants for each condition: (1) Y > N (*yes* rank / availability greater than *no* rank / availability); (2) Same (*equal* rank / availability for *yes* and *no*), and; (3) Y < N (*no* rank / availability greater than *yes* rank / availability). See Table 1 for details of the nine availability-ranking variations that we examined in our study.

Table 1. Variations in the availability and ranking of the yes-no content. Within each of the cells we show the availability on the first line and the rank on the second line.

		Availability		
Relative Rank		Y > N	Same	Y < N
	Y > N	More Yes Yes Higher	Same Yes-No Yes Higher	More No Yes Higher
	Same	More Yes Same Yes-No	Same Yes-No Same Yes-No	More No Same Yes-no
	Y < N	More Yes No Higher	Same Yes-No No Higher	More No No Higher

The result pages that we selected for this study had at least one *yes* and one *no* judgment in the top 10. We could vary the rank by ordering results related to one outcome above others, and randomly ordering them in the case of *same*. For availability, this was more challenging since we may need to introduce new results to offset initial imbalances in the result list. These results came from lower-ranked positions (primarily 11-20). The difference *d*, between the number of *yes* and *no* results for each of the queries ranged from 0-3; nine (25%) questions with each *d*. To do this, we found the minimum number of judgments for yes or no: *min*(*#yes,#no*). Then depending on whether we were favoring *yes* or *no* in the particular experiments we reduced the number of pages related to the non-preferred outcome by $d - min(\#yes,\#no)$ and increased the number of results for the preferred outcome by the same margin. For cases where *#yes* = *#no*, we dropped the lowest ranked *both* or *neither*-labeled caption and inserted a page with a *yes* or *no* judgment, depending on the preferred outcome. This process meant that we could automatically construct result pages with relative *yes-no* distributions to match the requirements specified in Table 1. The order of answer pages was initially randomized in the result pages, but held constant across all participants. We did

this to remove the effect of the original search engine ranking since we wanted to control the order experimentally. We inspected the random orderings to verify that *yes* and *no* were distributed evenly.

2.5 Summary

In this section we have described the procedures followed during the study. During the initial study we gathered 720 answer judgments before and after the review of the original (*unmodified*) rank ordering returned by the search engine (i.e., 36 questions and 20 judges per question). For the follow-up experiments we gathered an additional 6480 pre- and post-search answer judgments, over all manipulations in Table 1 (720 answer judgments for each of the nine cells, referred to as the *modified* ranking). The data collected as described in this section let us study belief dynamics during search in more detail than was possible in our prior log-based investigation (White, 2013), although in an artificial setting rather than a naturalistic setting as was possible with the log analysis. We discuss any implications of differences in the experimental settings later in the article.

3. Findings

We now present the findings of our analysis on the impact of searching on searcher beliefs. In Sections 3.1 and 3.2 we use the belief ratings provided before and after the review of search engine result lists. (modified and unmodified). In Section 3.3, we use the modified result lists arising from the various manipulations described earlier in Section 2.4 to examine their effect on belief dynamics. Section 3.4 discusses the accuracy of the answers that participants attained after reviewing the unmodified and modified ranked lists.

3.1 Belief Dynamics

Given that the overall percentage correctness did not change as a result of interacting with the engine, we wanted to better understand *how* beliefs changed during information search (e.g., were they tending more to *yes* than *no* or vice versa?). To do this we compared the pre- and post-search belief distributions. The distributions of responses across all questions and all participants are shown in Figures 3a and 3b. Also shown are pre- and post-search distributions for a modified result list with the *same-same* distribution (the middle cell in Table 1) which controls for variations in the rank and availability of answer content on the results page (see Figures 3c and 3d).

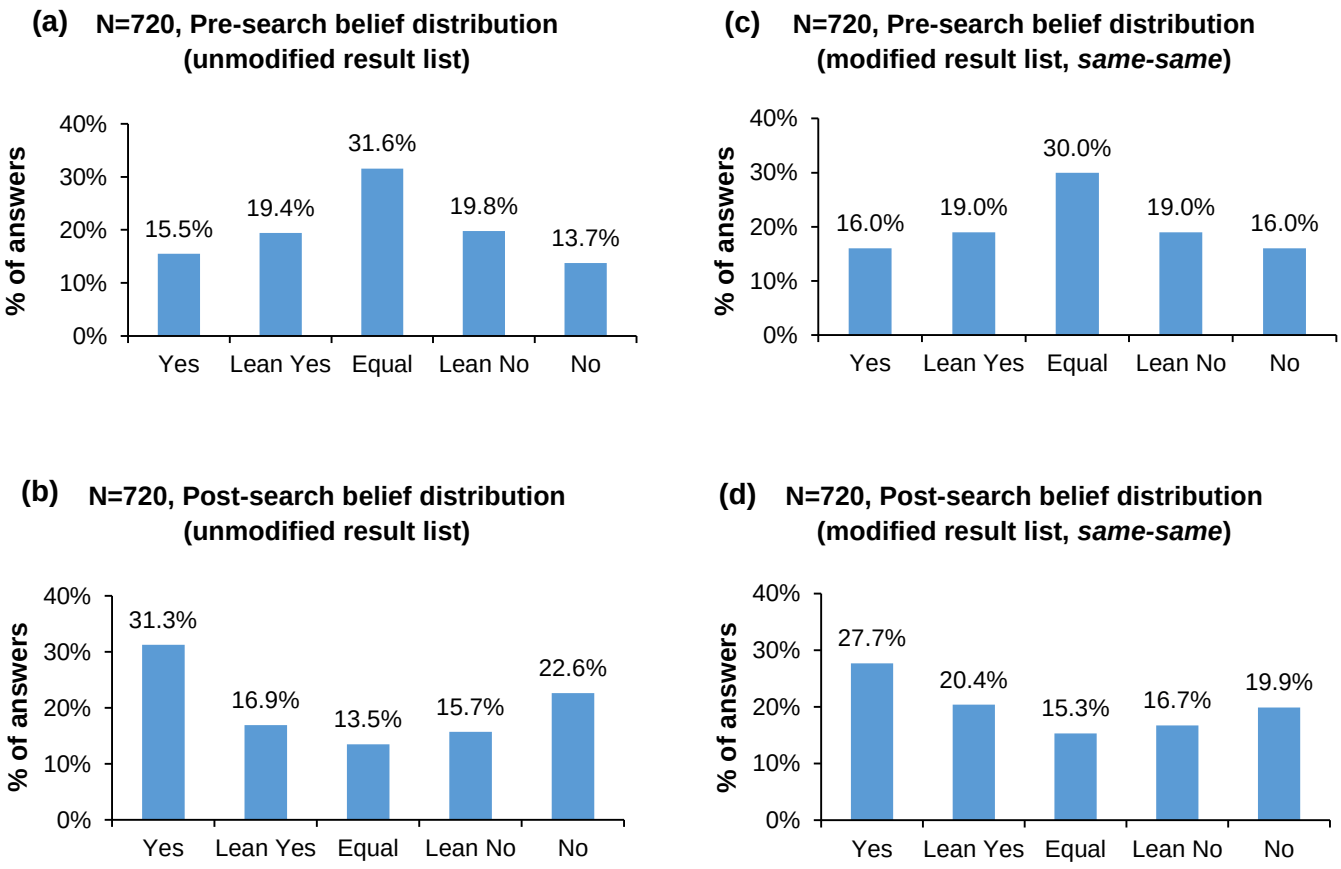

Figure 3. Outcome distributions provided by participants for (a) before reviewing search results (unmodified list), (b) after reviewing search results (unmodified list), (c) pre-search distribution (modified list, *same-same*), and (d) post-search distribution (modified list, *same-same*).

As expected, most participants were unsure at the outset of their search (31.6% in Figure 3a). The heightened uncertainty may relate to the assignment of potentially unfamiliar tasks to which people had no personal commitment. It is interesting to consider that in previous work (White, 2013), when we asked people to recall their pre-search beliefs from their own questions they were highly skewed toward *yes*, suggesting that confirmation was more likely than observed here. Although most participants were unsure (with an *equal* rating) there is still a sizeable fraction (approximately 30%) who strongly believed *yes* or *no* before they started searching. Reasons for this could include prior experience with the domain or the question, or personality traits that make them more confident in their belief assignment. We explore the relationship between these pre-search beliefs and search behaviors or outcomes later in the article.

Figure 3b shows the belief distribution after participants had reviewed the search results. As expected, we observe a reduction in uncertainty, with many fewer participants reporting that they were completely uncertain (*equal* drops from 31.6% to 13.5%), and over half of participants (54%) forming a strong opinion (*yes* or *no*) following review of the search results. Reductions in initial uncertainty following the review of

information is well acknowledged in information seeking research (Kuhlthau, 1991; Kuhlthau, 1993). One exception is Wilson et al. (2002, p. 709), who actually show an increase in uncertainty following searching. Most interesting is that participants were more likely to end up believing *yes* following examination of the result list than *no* (31.3% *yes* vs. 22.6% *no*), even though the correct answer is split evenly across all questions between these two outcomes. This mirrors the findings of White (2013), where we found through a retrospective survey that participants were more likely to believe *yes* following search.

There are at least two factors at play in the revision of such beliefs: (1) the availability of answer content in the results, and (2) searcher's prior beliefs and their impact on search behavior. The former explanation for the shift to *yes* is differences in answer availability in the results (e.g., more results answering *yes* than *no*), which may lead people to associate the greater availability of *yes*-oriented content with the greater likelihood of the *yes* outcome (Tversky and Kahneman, 1973). Indeed, upon analyzing the answer labels assigned to captions and results in the result pages of interest, we see that 44% of the results at position one are labeled *yes* (vs. 30% for *no*, remainder are *neither* or *both*), and 32% of the top-10 results are *yes* (vs. 24% for *no*).[2,3] However, looking at the distributions in Figure 3c and 3d we see that there is a similar trend in the findings (although not quite as clear), even though we control for rank and availability and there is not the same bias toward yes-related content as is observed in normal search engine operation. This suggests that the change in beliefs may relate to factors beyond availability, such as searchers' pre-search beliefs.

3.1.1 Direct Changes in Beliefs

We analyze the relationship between pre- and post-search beliefs in more detail. Figure 4a illustrates the fraction of pre-search beliefs that end up in each of the post-search outcomes given their value pre-search. Figure 4b shows the distribution of pre- and post-search beliefs for the modified (*same-same*) result list which balances rank/availability between *yes* and *no*, in a similar way to the previous analysis.

[2] In this work we focus on result pages with at least one *yes* and one *no* caption. In other research where we remove this restriction, we have seen similar biases toward *yes* information in result pages (White, 2013).
[3] This imbalance was attainable in this part of the study because we were using the original *unmodified* ranked list returned by the search engine for our analysis here.

(a) Unmodified

Pre-search \ *Post-search*	*Yes*	*Lean Yes*	*Equal*	*Lean No*	*No*
Yes	**68.40%**	8.90%	1.30%	6.30%	15.20%
Lean Yes	25.30%	26.30%	13.10%	21.20%	14.10%
Equal	20.50%	19.90%	**29.20%**	14.30%	16.10%
Lean No	18.80%	20.80%	6.90%	25.70%	27.70%
No	25.70%	7.10%	10.00%	12.90%	**44.30%**

(b) Modified

Pre-search \ *Post-search*	*Yes*	*Lean Yes*	*Equal*	*Lean No*	*No*
Yes	**55.40%**	16.10%	3.30%	4.50%	12.90%
Lean Yes	31.20%	26.10%	5.90%	20.50%	16.30%
Equal	22.90%	18.00%	**29.20%**	14.10%	15.80%
Lean No	20.70%	21.10%	7.90%	21.80%	28.60%
No	24.70%	13.50%	9.90%	14.10%	**37.80%**

Figure 4. (a) Post-search distribution of outcome likelihoods given pre-search belief (unmodified ranking), and (b) post-search distribution of outcome likelihoods given pre-search belief (modified ranking, *same-same*).

Figure 4a shows that participants were likely to retain their original belief (e.g., 68.4% of those who believe *yes* before search, believed *yes* after search), and more likely to transition to *yes* if there pre-search belief was *equal* (20.5% + 19.9% = 40.4% *yes* vs. 14.3% + 16.1% = 30.4% *no*). Over half (57.5%) of all shifts are toward *yes* and a quarter of participants who start believing *no* completely reversed to *yes* following the review of the search results. Participants were also more likely to retain their pre-search belief it was more strongly held (e.g., 44.3% of those who believed no beforehand retained that belief afterwards).

Interestingly, participants were more likely to retain a belief in *yes* than *no*. One explanation for this is that the search engine retrieved more *yes*-related content, reinforcing their beliefs. To understand whether this was the case, we repeated this analysis with the *same-same* dataset, which is not affected by such biases. The findings summarized in Figure 4b show that there is a similar effect on belief revision even when we control for rank and availability, although differences are slightly less pronounced than Figure 4a.

In both the sets of analysis we note that the changes in beliefs are typically minor (average=1.1 rating steps, median=1). Given that we know the correct answers to our set of yes-no questions, we can calculate that for 49.7% the questions answered, participants who had an outcome in mind (one of *yes*, *lean yes*, *lean no*, or *no*) had an incorrect pre-search belief. To transition from correct to incorrect requires a significant adjustment of belief (i.e., by more than two rating points given the five-point scale). However, most

revisions are slight, with only 16% of beliefs changing by this margin. Although there is some dynamism in searcher beliefs, the extent of the belief revision is insufficient to get searchers to correct answers.

3.2 Beliefs and Search Behavior

An important aspect of understanding why people do not dramatically update their beliefs during search is to study the nature of the search results that they select, in particular the answers residing in those results and their associated captions. To do this we used judgments associated with the result-page captions. As well as recording the pre- and post-search belief ratings, we also recorded participants' search interaction, in particular the number of results clicked and which results were clicked.

To ameliorate position and trust biases (as noted by Joachims et al. (2007)), we focused our analysis on the manipulation experiment where result pages had a balanced ranking and an equal availability of *yes* and *no* content. In that experiment, participants clicked at least one search result for 45.7% (n=329) of the 720 questions that they attempted to answer. For these participants we examine three behavioral features associated with click activity: (1) *clickthrough rate*: The fraction of result pages with any click on a search result; (2) *number of clicks:* The number of unique results clicked, and; (3) *nature of clicks:* The fraction of clicks that were on *yes*-only caption content and the fraction that were on *no*-only content for each of the pre-search beliefs. We conjectured that these features may differ between the groups, e.g., people with stronger beliefs may be more likely to seek confirmatory information. Table 2 summarizes the findings of this behavioral analysis for the modified list with the same-same rank-availability distribution.

Table 2. Search result engagement for modified result list (*same-same*). Percent or mean (standard deviation).

Measure	***Participant Group***					
	Pre=Yes	**Pre=Lean Yes**	**Pre=Equal**	**Pre=Lean No**	**Pre=No**	**All participants**
N	112	139	227	142	100	720
Clickthrough rate	33.9%	46.8%	52.9%	44.6%	42.0%	45.7%
# unique results clicked	1.65 (0.72)	2.03 (0.91)	2.43 (0.95)	1.85 (0.83)	1.89 (0.86)	1.94 (0.94)
% Click *yes* (vs. *no*)	68.2%	65.8%	57.3%	53.8%	37.0%	59.0%

Table 2 shows that participants who are unsure (Pre=Equal) were more likely to explore the search results, and more likely to click on multiple results than participants with stronger beliefs (*yes* or *no*). A one-way analyses of variance (ANOVA) shows that the difference in the number of clicks is significant ($F(4,311)=3.61$, $p < 0.01$), with *equal* and *yes* being significantly higher and lower respectively than the others per post-hoc testing ($p < 0.01$). Significant differences were also observed for the other measures (Chi-squared tests: both $\chi^2(4) \geq 12.27$, $p \leq 0.02$). Those who believed *no* at the outset were more likely to engage with the results than those believing *yes* (clickthrough rate = 42% for *no* vs. 34% for *yes*). It may be more difficult to verify *no* than *yes*. Previous research has convincingly argued that conclusively falsifying hypotheses is challenging, but is also central to scientific practice (Popper, 1959).

From the "All participants" column in Table 2 we can see that overall, irrespective of pre-search beliefs, people are more likely to click on *yes* than *no* and significantly more so than expected (50%) (test of proportions: Z=2.40, p=0.008). The percentage of people who click a caption with *yes*-only varies consistently as a function of the pre-search belief; there is a larger fraction of the clicks on *yes* for stronger pre-search belief in *yes*. Importantly, when the pre-search belief is *equal*, participants were still significantly more likely to click on *yes*-oriented results, even when we control for the availability and result ranking, and that this bias is still somewhat evident in the Pre=Lean No column (with 54% of result clicks on *yes*, although non-significantly (Z=0.57, p=0.29)). It is not until people strongly believe *no* that we observe any preference for *no*-related captions (63% of clicks on *no*-only captions, Z=2.44, p=0.01).

While we have shown that pre-search beliefs influence click behavior, the analysis presented in this section so far is insufficient to determine whether the post-search answer relates solely to the results clicked or also includes other factors such as pre-search belief. To help answer this question, we computed the point biserial correlation coefficient (r_{pb}) between the percentage of clicks on pages with *yes* and whether they selected the *yes* answer (binary). Table 3 reports the percentage of answers that are *yes* from each group as well as the correlation between the percentage of clicks on *yes*-related pages and the answer chosen for each of the participant groups.

Table 3. Percentage of answers provided that are *yes* and the correlation with percentage clicks on *yes* (r_{pb}) for each participant group. Correlations significant at ** $p < 0.01$ and *** $p < 0.001$.

Measure	***Participant Group***					
	Pre=Yes	**Pre=Lean Yes**	**Pre=Equal**	**Pre=Lean No**	**Pre=No**	**All participants**
% Answer *yes* (vs. *no*)	71.4%	67.3%	59.9%	44.7%	40.2%	58.6%
r_{pb} with % clicks on *yes*	0.332***	0.337***	0.429***	0.222**	0.207**	0.354***

Table 3 shows that the correlations between the percentage of answers that are *yes* and the number of clicks of *yes* are around 0.3, peaking at Pre=Equal. This suggests that the participants who are least sure of their answer (Pre=Equal) relied most on the results viewed. The answers of those who believed *no* were also less correlated with the fraction of *yes* clicks than those who believed *yes*, suggesting that that the *no* groups were less likely to be affected by exposure to *yes*-related information. If r_{pb}= 1 then the *yes* answer percentages could be fully explained by variations in the fraction of result clicks with *yes* answers. However, since this is not the case (in fact quite far from it), we can conclude from these findings that both the results visited *and* searcher's prior beliefs (and perhaps other unmodeled factors) contribute to decisions regarding the answers that participants eventually settle on.

While search engines cannot affect searchers' prior beliefs, they can influence the results that they click on by using different presentation strategies and they can consider searchers' prior beliefs through mechanisms such as personalization (e.g., searchers with weaker beliefs may be more open to considering

alternative outcomes, as evidenced by their willingness to change beliefs from Pre=Equal in Figure 4a and 4b). We wanted to understand whether we could shift pre-search beliefs (including strongly-held beliefs) by manipulating the search results, with a view toward ultimately directing searchers to the correct answers by integrating factual correctness in search engine ranking. We also wanted to understand participants' receptiveness to answer content, not just overall, but also as a function of differences in pre-search beliefs.

3.3 Manipulating Beliefs

Since people inspect result lists from top-to-bottom and typically trust the engine ranking (Joachims et al., 2007), one way that a search engine could affect belief revision is to integrate consensus or base-rate information into its ranking algorithm, and compute a query-dependent correctness score for each result, that includes factors such as the reputability of the source as well as the accuracy of the answer contained. In doing so, they could exclude incorrect information completely from the ranked list or rank results from correct to incorrect to provide room for error in the accuracy estimates. To collect factual information, search engines could mine knowledge repositories such as the medical literature or Web content (Dumais et al., 2002; Ferrucci et al., 2010), or apply human computation methods (Von Ahn et al., 2006), ultimately combining their output with other ranking features to order or filter results.

In addition to studying the effect of the original (unmodified) search engine ranking on beliefs, we examined the impact of two manipulations of the search results, three variants of each: (1) availability of captions/results with *yes* or *no* (e.g., less *yes* than *no*; Y < N), and (2) the relative ordering of *yes* and *no* answer pages in the result list (e.g., all *yes*, then all *no*; Y > N). The full details of these manipulations were provided earlier in Section 2.4. We focus on the extent to which people's beliefs change as a function of applying these manipulations. This required running the same experiment as above, but for each of the nine variants of availability and ranking in Table 1. A total of 720 answers (36 questions × 20 answers to each question) were collected for each of the nine experiments, 6480 (9 × 720) answers in total.

The pre-search beliefs for each of the nine conditions were distributed similarly to that in Figures 3a and 3c with little variation (for space reasons we do not present those here). Table 3 shows how those beliefs changed for all participants and for the five pre-search groups (*yes*, *lean yes*, etc.). For simplicity and since we focus on trends rather than exact percentages, we use sparklines to denote belief distributions from *yes* (left) to *no* (right). The sparklines are compressed versions of Figure 3b (post-search belief distribution), with *y*-max=84.7% for the highest fraction of participants with a belief (when pre = *yes*, availability = Y > N, rank = Y > N). To quantify the variation in belief distributions, we compute the Fisher-Pearson coefficient of skewness (Doane and Seward, 2011) for each of the distributions; a positive value quantifies skew toward *yes* and a negative value represents the skew toward *no*. The skewness value is shown above each sparkline in Table 4 along with indications of statistical significance, defined in the table caption.

Table 4. Post-search beliefs for all participants and the manipulations of availability and order (*y*-max = 84.7%). Skewness and statistical significance also shown* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Participant Group		Availability		
All participants		Y > N	Same	Y < N
Rank	Y > N	0.825***	0.366***	0.176
	Same	0.486***	0.391**	0.137
	Y < N	0.015	−0.319**	−0.486**
Pre = Yes		Y > N	Same	Y < N
Rank	Y > N	3.254***	1.633***	1.479***
	Same	2.286***	1.395***	0.865***
	Y < N	1.388***	0.382	0.273
Pre = Lean Yes		Y > N	Same	Y < N
Rank	Y > N	1.870***	0.845**	0.598*
	Same	1.076***	0.670**	0.530*
	Y < N	0.154	−0.014	−0.047
Pre = Equal		Y > N	Same	Y < N
Rank	Y > N	0.846***	0.435*	0.029
	Same	0.362*	0.179	0.085
	Y < N	−0.083	−0.215	−0.577**
Pre = Lean No		Y > N	Same	Y < N
Rank	Y > N	0.382	−0.057	−0.321
	Same	0.040	0.022	−0.107
	Y < N	−0.454	−1.014***	−1.218***
Pre = No		Y > N	Same	Y < N
Rank	Y > N	−0.015	−0.338	−0.449
	Same	−0.181	−0.714**	−0.867***
	Y < N	−0.445	−0.951***	−1.168***

Table 4 ("All participants") shows that overall the post-search beliefs can be affected by the volume and ranking of content pertaining to certain outcomes. In particular, we draw reader attention to the cells in the upper left (availability: Y > N, rank: Y > N) and the lower right (availability: Y < N, rank: Y < N) (both above the double line toward the top of Table 3), which show that overall across all searchers, post-search beliefs can be strongly influenced by the availability and rank ordering of answers in search results. We can clearly observe that the belief distributions reflected in those cells largely match the distribution of the underlying results. It appears possible that search engines could help searchers update beliefs to find the correct answer (if one is exists and is known to the search engine).

Table 4 also shows that those with weaker beliefs (*lean yes*, *equal*, *lean no*) are more likely to be affected by result manipulations. As we show in the next section, this can have a significant impact on search success. Importantly, participants who were confident initially seldom revised their beliefs, even in light of strong contradictory evidence. This aligns other recent studies on belief updating in information access (Liao and Fu, 2013), although our focus is on search engine results rather than side-by-side

comparisons, and we target questions with definitive answers rather than controversial topics. Our focus on answerable questions allows us to study the effect of these strongly-held beliefs on factual correctness. We explore answer accuracy in more detail in the next section.

3.4 Answer Accuracy

We performed two analyses of answer accuracy: (1) the effect of the original search engine ranking, and (2) the effect of the result manipulations described in the previous section. In performing this analysis we consider the *yes* and *lean yes* responses as a single *yes* group and the *no* and *lean no* responses as a single *no* group. Grouping the answers in this way simplified the computation of answer accuracy.

3.4.1 Effect of Original Engine Ranking

In the first part of this analysis, we explore the effect of the original (unaltered) search engine results on the accuracy of the results that were obtained. As noted earlier, prior to using the search engine participants had the correct answer for 50.3% of questions (if we exclude participants from Pre=Equal). Following review of the search results presented, the answer accuracy was found to be little changed at 53.6% (again excluding those from Pre=Equal). Although there was a slight increase in answer accuracy as a result of searching, it appears that on average the search engine did not provide much assistance in improving the accuracy of the answers that our participants attained.

From analyzing the answers assigned to captions and results, we see that the correctness of the search engine's top result, when it has only *yes* or only *no* content (74% of queries), is near random: 53% for both sources. Figure 1b shows an example of this for the question query "does mono in children cause bruising?" where the top results lean toward the incorrect answer (*yes*). The top-ranked search results are particularly important since gaze tracking studies have shown that they are most frequently clicked on by searchers (Joachims et al., 2007). By considering factual correctness in ranking, search engines could promote correct answers and help searchers find correct answers. However, the effectiveness of such an approach is dependent on searchers following the direction of the search engine. This also depends on beliefs being impacted by the answers on clicked pages, something that we showed earlier (in Table 3) varies by both pre-search belief strength and pre-search belief outcome. In the next section we explore whether this is plausible using a subset of the questions where the result manipulation from Table 1 agreed with the consensus answer from physicians.

3.4.2 Effect of Result Manipulations

Within each of the cells in Table 2 where the availability and ranking point in the same direction (i.e., both Y > N or both Y < N), we selected the 18 questions where there correct answer agreed with the manipulation (e.g., correct answer = *yes*, Y > N for both availability and rank). We measured accuracy across all participants and for those with each of the pre-search beliefs. Table 5 shows the accuracy statistics.

Table 5. Answer correctness per participant group.
Confirmation = correct answer concurs with pre-search belief.
Disconfirmation = correct answer refutes pre-search belief.

Participant Group	**Answer = *yes* and Y > N**		**Answer = *no* and Y < N**	
All participants	74.9%		63.1%	
Pre = Yes	Confirmation	90.1%	Disconfirmation	38.9%
Pre = Lean Yes		83.1%		54.1%
Pre = Equal		70.7%		66.7%
Pre = Lean No		68.7%		71.0%
Pre = No	Disconfirmation	44.4%	Confirmation	87.0%

The findings for "All participants" in Table 5 suggest that on average there was a significant improvement in answer accuracy from adhering to modified result ordering provided by the search engine. If the answer was *yes*, accuracy grew from 55.3% to 74.9% ($Z = 5.30$, $p < 0.001$) and if the answer was *no* accuracy improved from 45.7% to 63.1% ($Z = 4.64$, $p < 0.001$). This demonstrates that on average searchers were amenable to changing their beliefs and that their answer accuracy could be improved if they followed the search engine, when it was correct.

We also observed that participants' pre-search beliefs greatly influenced their post-search success. Table 5 shows that those participants with strong beliefs initially and who resisted the strong evidence refuting their beliefs frequently obtained the incorrect answer (in around 60% of the observed answers, shaded gray cells in Table 5). These participants may be facing cognitive dissonance associated with an unpleasant state of post-decisional conflict brought on by attitude-challenging information, leading them to either avoid information that is inconsistent with their beliefs or find ways to refute that contradictory information (Festinger, 1957; Frey 1986). Perhaps as a consequence, we observe participants selecting results that support their belief significantly more than otherwise (i.e., %*yes* given Pre=Yes = 63.5%, %*no* given Pre=No = 60.1%; both $Z \geq 2.38$, $p \leq 0.01$). Even when we attempt to direct participants to the correct answer, they still selectively examined results confirming their prior beliefs. Those without strong pre-search beliefs followed the search engine in these cases and therefore achieved better accuracy. Realizing these gains is dependent on search engines knowing and using the correct answer, something that results in Section 3.4.1 suggest they are do not do at present. This is an interesting and important area for future work.

4. Discussion and Implications

Our user study allows to examine belief dynamics during Web search. Our findings showed that pre-search beliefs are affected only slightly by searching and any changes in belief are more likely to move toward positive (*yes*). Even when we controlled for engine ranking effects, participants were observed clicking on *yes* content more than *no* content (around 60/40); in the aggregate, this has the potential to affect machine-learned rankings. Pre-search beliefs were also shown to affect search behavior and strongly-held beliefs are difficult to shift (at least using variations in the availability and ranking of search results – the primary

mechanisms available to search engines), and this can affect answer accuracy. Search engines may exhibit near-random answer accuracy for yes-no questions and can shift some beliefs by manipulating the relative ordering and availability of answers in the results. Search engines are not designed to answer questions—even though searchers may expect them to be—and the answer pages surfaced may be determined based only on rudimentary term matches or affected by biases from aggregated behavioral data, not reality as is needed to fulfill their function. Finally, we showed that strongly-held beliefs are both difficult to move using uncongenial information and can be counterproductive. Nonetheless, many searchers may be amenable to changing their beliefs and the search engine may be able to offer guidance in doing so by varying the availability and ordering of the answer pages in the search results.

The role of prior beliefs in shaping search interaction has implications for search system design because search engines do not consider the factual correctness, and learn from search behavior but currently ignore whether observed behaviors are driven by cognitive biases such as confirming prior beliefs, or testing hypotheses with particular strategies. Topic-related biases may also apply in personalized search settings where the search engine may choose to exclude certain information in light of a user's recent search interests (Pariser, 2011). By considering the cognitive factors affecting result selection decisions as well as the click evidence itself (e.g., whether clicks are associated with biased beliefs or escalatory content (White and Horvitz, 2013)), more accurate ranking models can be developed. Overall, methods for better understanding and detecting bias-related search interaction are required, e.g., for question queries, clicks on captions with supporting (*yes*-oriented) evidence could be down-weighted since we observe those to occur significantly more frequently than expected (59%) given base rates (50%) and those clicks may be driven by factors beyond relevance.

The process by which queries are generated and transition from a visceral to compromised state has been explored extensively in information retrieval research (Taylor, 1968), but we also need to better understand how people frame their beliefs in search queries (Levin et al., 1998; Tversky and Kahneman, 1981). For example, an automated analysis of the search engine logs from White (2013) revealed that only a small fraction (< 0.1%) of yes-no question queries were phrased negatively (e.g., "is congestive heart failure not a heart attack?"). The impact of this positive framing and other factors on how search engines interpret queries and rank pages with direct answers needs to be clearly understood before improvements can be made in the ranking process.

The findings presented in this study mirror those from our earlier research and complement the log-based methodology employed in that investigation (White, 2013). The log-based study was performed retrospectively with search engine usage data gathered from consenting Microsoft Bing searchers. Although the user study enabled us to capture additional insights about belief dynamics that were not available in the

earlier log analysis, one drawback of the current experiment is that the original search queries came from searchers who almost certainly were highly motivated to seek information important to them based on their concerns about health (probably themselves or a family member). The participants in the study described in this article were likely concerned about or motivated to learn information about few if any of the questions. However, despite these potential differences in the underlying motivations for searching, many of the results from the log-based study were replicated and extended in the user study presented in this article. For example, in our earlier log analysis as in this user study, we also estimated that searchers reached the incorrect answer around half of the time when using the search engine and that searchers were strongly drawn toward supporting (primarily *yes*-oriented) information.

We focus on a manageable set of questions for our judges (30 minutes of judging), and on the medical domain given its importance and the availability of answers. We restricted our analysis to the health domain, primarily because of the availability of consensus answers from physicians, but also because of the significance of the domain and the impact of content encountered on perceptions and healthcare utilization (White and Horvitz, 2010). However, further work is required to understand whether these findings hold in other domains where search interactions may be less likely to be driven by anxiety or concerns. With this in mind, we are reassured by the findings described in the previous paragraph, where we observed similar findings for people with limited emotional attachment to the task (the third-party judges) as with those who may be directly connected to it (the logged Bing searchers). We focused on Microsoft Bing in this analysis to be consistent with our earlier log-based study (White 2013). However, follow-up studies are necessary with other search engines before any claims can be made about the generalizability of our findings.

Finally, we showed that searchers with strong beliefs were less affected by search result ranking (even when presented with strong contradictory evidence) and that retaining these beliefs may be counterproductive. Promoting confirmatory information facilitates psychological stability and personal validation, promoting accurate information promotes accurate perceptions of reality. If search engines are to help people attain accurate answers, steps need to be taken to factor factual correctness into ranking algorithms, and more directly influence beliefs (e.g., with prominent inline answers), rather than relying on changes to answer availability and rank ordering. Persuasive methods could also be applied to convince users to shift beliefs (Johnson and Eagly, 1989; Petty and Cacioppo 1990) or reflect on them (Kriplean et al., 2012). The tension between actual correctness and feelings of validation has been studied in psychology (Hart et al., 2009), but this research needs to be integrated into search systems. Techniques to do this include uncovering correct answers using external resources or modeling of searchers' cognitive mechanisms to build more complete personalized search models that can find supportive results or weight clickthrough evidence from these users differently based on inferences about underlying motivations, beliefs, or biases. The relative importance assigned to correctness versus validation is also topic dependent: for consequential

topics such as health, search engines may favor accuracy, especially if there is a known answer; whereas for controversial topics such as politics or religion, search engines may favor diversity or belief validation since there may be no definitive answer or searchers' beliefs may be difficult to change (Liao and Fu, 2013).

5. Conclusions and Future Work

In this article, we have studied belief dynamics during Web search via a user study of yes-no questions in the medical domain. Our results show that even for assigned questions where participants may not have a personal commitment, they often stick with the original belief and retain that belief as they selectively explore the results presented by the search engine. Changes in belief are typically small as a result of searching, and remained that way for some searchers even when we manipulate the relative availability and ordering of answers in results to isolate the effects of pre-search beliefs. Participants who had strong pre-search beliefs related to a particular outcome were unlikely to change those beliefs even in light of significant contradictory evidence when we manipulated the result list. In cases where searchers had a strong incorrect belief and were presented with contradictory information through our result manipulations, only around 40% attained the correct answer (see Table 5). In contrast, those who were less certain initially were significantly more likely to alter their beliefs and were more likely (around 70%) to obtain the correct answer. There is a particular opportunity to help these receptive searchers if retrieval systems became aware of the correct answer and could model likely receptiveness to information content as part of personalization. Search engines need to consider beliefs and biases, including better modeling base rates regarding outcome likelihoods, considering both answer accuracy and belief validation in ranking, encoding belief estimates rather than topical interests and preferences in personalized search models, and integrating more targeted alerts to help encourage searchers to correct strongly-held but factually-incorrect beliefs. Among key future directions in this area is the development and validation of more sophisticated, theoretically-guided models of the relationship between factors such as prior beliefs, result clicks, and searcher belief revision.

References

Agichtein, E., Brill, E., and Dumais, S. (2006). Improving web search ranking by incorporating user behavior information. *Proc. SIGIR*, 19−26.

Anderson, N. (1981). *Foundations of Information Integration Theory*. New York: Academic Press.

Ariely, D. (2008). *Predictably Irrational: The Hidden Forces that Shape Our Decisions*. Harper Collins.

Baron, J. (2007). *Thinking and Deciding*. Cambridge University Press.

Bates, M.J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5): 407−424.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53: 370−418.

Belkin, N.J., Oddy, R.N., and Brooks, H.M. (1982). ASK for information retrieval: Part I. Background and theory. *Journal of Documentation*, 38(2): 61−71.

Brookes, B.C. (1980). The foundations of information science. Part I. Philosophical aspects. *Journal of Information Science*, 2: 125−133.

Cho, J. and Roy, S. (2004). Impact of search engines on page popularity. *Proc. WWW*, 20–29.

Dervin, B. (1983). An overview of sense-making research: concepts, methods and results. *Proc. Annual Meeting of the International Communication Association*.

Doane, D.P. and Seward, L.E. (2011). Measuring skewness: a forgotten statistic? *Journal of Statistics Education*, 19(2).

Dumais, S., Banko, M., Brill, E., Lin, J., and Ng, A. (2002). Web question answering: is more always better? *Proc. SIGIR*, 291–298.

Eagly, A.H. and Chaiken, S. (1993). *The Psychology of Attitudes.* Fort Worth, TX: Harcourt Brace Jovanovich.

Elster, J. (1979). *Ulysses and the Sirens: Studies in Rationality and Irrationality*. Cambridge University Press.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., and Welty, C. (2010). Building Watson: an overview of the DeepQA project. *AI Magazine*, 31(3): 59–79.

Festinger, L. (1957). *A Theory of Cognitive Dissonance.* Stanford, CA: Stanford University Press.

Fischer, P., Kastenmuller, A., Greitemeyer, T., Fischer, J., and Frey, D. (2011). Threat and selective exposure: the moderating role of threat and decision context on confirmatory information search after decision. *Journal of Experimental Psychology: General*, 140(1): 51–62.

Fortunato, S., Flammini, A., Menczer, F., and Vespignani, A. (2006). Topical interests and the mitigation of search engine bias. *Proceedings of the National Academy of Sciences*, 103(34): 12684–12689.

Frey, D. (1986). Recent research on selective exposure to information. *Advances in Experimental Social Psychology*, 19: 41–80.

Gigerenzer, G. and Todd, P.M. (2000). *Simple Heuristics That Make Us Smart*. Oxford University Press.

Goldman, E. (2006). Search engine bias and the demise of search utopianism. *Yale Journal of Law and Technology*, 188.

Hart, W., Albarracin, D., Eagly, A., Brechan, I., and Lindberg, M. (2009). Feeling validated versus being correct: a meta-analysis of selective exposure to information. *Psychological Bulletin*, 135(4): 555–588.

Hogarth, R.M. and Einhorn, H.J. (1992). Order effects in belief updating: the belief-adjustment model. *Cognitive Psychology*, 24: 1–55.

Ieong, S., Mishra, N., Sadikov, E., and Zhang, I. (2012). Domain bias in web search. *Proc. WSDM*, 413–422.

Ingwersen, P. (1994). Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. *Proc. SIGIR*, 101–110.

Joachims, T. (2002). Optimizing search engines using clickthrough data. *Proc. SIGKDD*, 133–142.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., and Gay, G. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems*, 25(2).

Johnson, B.J, and Eagly, A. (1989). Effects of involvement on persuasion: a meta-analysis. *Psychological Bulletin*, 106(2): 290–314.

Kimble, C. (2013). Knowledge management, codification and tacit knowledge. *Information Research*, 18(2), paper 577. [Available at http://informationr.net/ir/18-2/paper577.html].

Klayman, J. and Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94: 211–228.

Kuhlthau, C.C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science and Technology*, 42(5): 361–371.

Kuhlthau, C.C. (1993). A principle of uncertainty for information seeking. *Journal of Documentation*, 49 (4): 339-355.

Kriplean, T., Morgan, J., Freelon, D., Borning, A., and Bennett, A. (2012). Supporting reflective public thought with considerit. *Proc. CSCW*, 265–274.

Lauckner, C and Hsieh, G. (2013). The presentation of health-related search results and its impact on negative emotional outcomes. *Proc. SIGCHI*, 333–342.

Levin, I.P., Schneider, S.L., and Gaeth, G.J. (1998). All frames are not created equal: a typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes*, 76(2): 149–188.

Liao, Q. and Fu, W.T. (2013). Beyond the filter bubble: Interactive effects of perceived threat and topic involvement on selective exposure to information. *Proc. SIGCHI*, 2359–2368.

Mankoff, J., Kuksenok, K., Rode, J., Kiesler, S., and Waldman, K. (2011). Competing online viewpoints and models of chronic illness. *Proc. SIGCHI*, 589–597.

Marchionini, G. (1995). *Information Seeking in Electronic Environments*. Cambridge University Press.

Munson, S. and Resnick, P. (2010). Presenting diverse political opinions: how and how much. *Proc. SIGCHI*, 1457–1466.

Pariser, E. (2011). *The Filter Bubble: What is the Internet Hiding from You?* Penguin Press.

Petty, R.E. and Cacioppo, J. (1990). Involvement and persuasion: tradition versus integration. *Psychological Bulletin*, 107(3): 367–374.

Pitkow, J., Schütze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., Adar, E., and Breuel, T. (2002). Personalized search. *Communications of the ACM*, 45(9): 50–55.

Popper, K. (1959). *The Logic of Scientific Discovery*. Mohr Siebeck.

Rochman, B. (2011). Jenny McCarthy, vaccine expert? A quarter of parents trust celebrities. *Time*. Published 26 April 2011. Retrieved 6 May 2013.

Rose, D. and Levinson, D. (2004). Understanding user goals in Web search. *Proc. WWW*, 13–19.

Simon, H. (1991). Bounded rationality and organizational learning. *Organization Science*, 2(1): 125–134.

Taylor, R.S. (1968). Question-negotiation and information seeking in libraries. *College and Research Libraries*, 29: 178–194.

Teevan, J., Dumais, S.T., and Horvitz, E. (2005). Personalizing search via automated analysis of interests and activities. *Proc. SIGIR*, 449–456.

Tversky, A. and Kahneman, D. (1973). Availability: a heuristic for judging frequency and probability. *Cognitive Psychology*, 5(1): 207–233.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, 185: 1124–1130.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4811): 453–458.

Von Ahn, L., Kedia, M., and Blum, M. (2006). Verbosity: a game for collecting common-sense facts. *Proc. SIGCHI*, 75–78.

Wason, P.C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12: 129–140.

White, R.W. (2013). Beliefs and biases in web search. *Proc. SIGIR*, 3–12.

White, R.W. and Horvitz, E. (2009). Cyberchondria: studies of the escalation of medical concerns in web search. *ACM Transactions on Information Systems*, 27(4): 23.

White R.W. and Horvitz, E. (2010). Web to world: predicting transitions from self-diagnosis to the pursuit of local medical assistance in web search. *Proc. AMIA*, 882−886.

White R.W. and Horvitz, E. (in press). Captions and biases in diagnostic search. *ACM Transactions on the Web*.

Wilson, T.D., Ford, N., Ellis, D., Foster, A., and Spink, A. (2002). Information seeking and mediated searching: Part 2. Uncertainty and its correlates. *JASIST*, 53(9): 704–715.

Yue, Y., Patel, R., and Roehrig, H. (2010). Beyond position bias: examining result attractiveness as a source of presentation bias in click-through data. *Proc. WWW*, 1011–1018.