Beliefs and Biases in Web Search

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

ABSTRACT

People's beliefs, and unconscious biases that arise from those beliefs, influence their judgment, decision making, and actions, as is commonly accepted among psychologists. Biases can be observed in information retrieval in situations where searchers seek or are presented with information that significantly deviates from the truth. There is little understanding of the impact of such biases in search. In this paper we study search-related biases via multiple probes: an exploratory retrospective survey, human labeling of the captions and results returned by a Web search engine, and a large-scale log analysis of search behavior on that engine. Targeting yes-no questions in the critical domain of health search, we show that Web searchers exhibit their own biases and are also subject to bias from the search engine. We clearly observe searchers favoring positive information over negative and more than expected given base rates based on consensus answers from physicians. We also show that search engines strongly favor a particular, usually positive, perspective, irrespective of the truth. Importantly, we show that these biases can be counterproductive and affect search outcomes; in our study, around half of the answers that searchers settled on were actually incorrect. Our findings have implications for search engine design, including the development of ranking algorithms that consider the desire to satisfy searchers (by validating their beliefs) and providing accurate answers and properly considering base rates. Incorporating likelihood information into search is particularly important for consequential tasks, such as those with a medical focus.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Search process, Selection process.*

Keywords

Beliefs; Biases; Search interaction; Health search.

1. INTRODUCTION

Information scientists have analyzed the cognitive mechanisms behind the search for information, including the development of models for how information needs emerge [4][37] and how they evolve during search [23][24]. The models developed typically focus on a cognitive actor, and that actor's interactions with information objects and systems, within a context [17][30]. Although they may model historic search interests, these models do not consider prior *beliefs* about outcomes and associated likelihoods, or biases arising from those beliefs. Biases in cognition may lead people to create beliefs based on false premises and behave in a seemingly irrational manner [21], e.g., exhibiting preference for information in support of their position over information refuting it, irrespective of factual correctness [2][3][21]. For example, although there is a definite answer to the yes-no question *Can tea tree oil treat canker sores?*, and that answer is *yes*, a health seeker may favor a particular outcome in light of their beliefs about the value of the oil, and seek or unconsciously prefer disaffirming information.

Biases can be observed in information retrieval (IR) in situations *where searchers seek or are presented with information that significantly deviates from true likelihoods*. In IR, the term "bias" has been associated with search engine functionality (e.g., caption generation [47]) and user preferences (for higher ranked results [20] or particular domains [16]), but not seemingly irrational behaviors or skewed result lists as we focus on here. Although the motivation between previous studies and ours differs, the impact of the biases can be similar: both types cause searchers to be interested in particular results for reasons beyond relevance. As such, biased beliefs can also affect aggregated behavioral signals (e.g., result clickthrough) used by search engines for ranking [1][19]. As we show here, when an answer to a yes-no medical question is provided in the top result from a search engine, more than half of the time that answer is incorrect. Search engines need to consider both satisfying users (by surfacing results that reinforce biases, but could also be factually incorrect) and providing correct answers (e.g., performing bias mitigation to provide more rational result sets that accurately reflect background probabilities). For important searches such as those in the health domain—where people have been shown to interpret the result ranking as an ordering of condition likelihoods [44]—search engines may wish to favor accuracy over satisfaction given the potentially-serious ramifications of an incorrect answer.

In this paper we present the first comprehensive study of beliefs and biases in search, highlighting the potentially counterproductive effects of biases on answer accuracy. Focused on the health domain, we demonstrate the large extent to which bias is evident in search behaviors and results, and we argue that biases need to be considered by the IR community in ranking and interaction support. To study these biases we employ a range of methods, including an initial exploratory survey to gain insight into prior beliefs and how they change as a result of search, human labeling (by crowdsourced judges as well as physicians), and a log analysis of Web search behavior. We study biases in the context of yes-no questions because we can attain direct ground truth answers via expert consensus, allowing us to study answer accuracy, central to our definition of bias. As we show, yes-no questions are also popular (2% of the queries in our search engine log sample were yes-no questions) and they offer an excellent opportunity for search engines to help users since there is often a likely answer. This study answers the following research questions: (i) Do people's beliefs in different outcomes (*yes* versus *no* in this case) change (and how) as a result of searching? (ii) To what extent are search engine results biased in favor of particular outcomes? (iii) To what extent do biases appear to affect search behaviors? and (iv) What is the impact of these factors on search outcomes, specifically answer accuracy? Answers to these questions help us better understand the role of biases in search

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

and can inform the design of IR methods to consider biases, e.g., in deciding when accuracy should feature in result ranking, or when systems should diversify the result set to expose new perspectives.

The remainder of the paper is structured as follows. Section 2 describes related work in the psychology and IR research communities. Section 3 presents the results of a large exploratory survey of biases in beliefs during the search process. Section 4 describes the identification of yes-no question content labels and definitive answers necessary for our analysis. Section 5 describes the extent to which search engine result pages (SERPs) and the result ranking is skewed toward a particular answer for the yes-no questions we select. Section 6 focuses on search behavior, including the accuracy of the answers that searchers appeared to find. We discuss the findings and their implications in Section 7 and conclude in Section 8.

2. RELATED WORK

Research in a number of areas is relevant to the work described here, namely: (i) cognitive biases and decision making, (ii) utilizing search behavior to rank results, (iii) biases in search interaction, and (iv) personalization and diversity. We consider each area in turn.

Numerous models of the search process have been developed over the past few decades [4][17][23][24][30]. These models primarily focus on information need formulation, search interaction, resolving uncertainty, and contextual influences on the search process. However, these models largely ignore situations when seemingly irrational search behavior is observed, e.g., people accepting factually incorrect or unsupported information because it reinforces a particular belief they hold. Cognitive biases are defined as a pattern of deviation in judgment occurring in particular situations, where a deviation may refer to a difference from what is normatively expected, either by the judgment of people outside of the situation or by independently verifiable facts [21]. Biases play a central role in human judgment and decision making and span a number of different dimensions (see [2][3][21] for summaries). Other models of irrational human behavior are plentiful, including related concepts such as bounded rationality [33]. Biases can be difficult to distinguish and are not necessarily negative. For example, they can form information-processing shortcuts leading to more effective actions in a given context or enable faster decisions when timeliness is more valuable than accuracy, e.g., the availability heuristic allows rapid assessments of likelihoods based on ease of recall [15][21].

Confirmation bias describes people's unconscious tendency to prefer confirmatory information [26][41]. This can make searchers more likely to employ positive test strategies, where people seek evidence that supports their hypothesis and disregard evidence that refutes it [22]. Moving beyond behavior, biases in the *results* returned toward one perspective can reinforce existing beliefs and leave searchers susceptible to the effects of information availability [33][39], where people's ability to make rational decisions is limited by the information that they have access to [33] and the ease with which information is recalled (via the engine in this case) maps to likelihood estimates [39]. Although there has been extensive research on cognitive processes in search, we are the first to present a detailed investigation of beliefs and biases in search. The role of biases in search were discussed in previous studies on "cyberchondria" [44], but were not studied as directly as we do in this paper.

Over the past decade, a number of authors have proposed methods for using behavioral data of various forms—including queries, result clicks, and post-click navigation behavior—to improve result relevance [1][4][19]. Despite its utility for ranking, search behavior can still be affected by biases related to the ordering or presentation of results on the SERP, or user preferences for particular resources. Joachims et al. [20] analyzed searchers' decision processes via gaze tracking and compared implicit feedback from search-result clicks against manual relevance judgments. They found that clicks are informative but biased (favoring results at higher rank positions), yet relative result preferences derived from clicks mirror searchers' true preferences. Searcher models can capitalize on this consistent behavior to infer search result attractiveness and document relevance (e.g., [10]). Over time, searchers' focus of attention on top-ranked content can create a vicious cycle whereby clicks reinforce popular results [8], although this popularity bias may be offset somewhat by the heterogeneity of searchers' topical interests [14].

Other factors beyond rank position can introduce bias into search. Clarke and colleagues [9] introduced click inversions to study features of the captions that increase caption attractiveness. Yue and colleagues [47] studied the effect of caption attractiveness, defined for their study as the presence and absence of bolded terms in the titles and snippets of the caption. They show via experiments conducted on the Google Web search engine substantial evidence of presentation bias in clicks towards results with more attractive titles. Beyond captions, Ieong and colleagues [16] studied the effect of domain biases, whereby a result is believed to be more relevant because of its source domain. They show that this bias exists in click behaviors as well as human judgments, and that domain can flip caption preferences around a quarter of the time, independent of rank or relevance. However, these biases are unrelated to biased beliefs about task outcomes and the role of search engines in reinforcing those beliefs, as we target in this study.

Beyond aggregating search behavior across all users, search engines can also cater to the individual needs of their users via personalization, presenting the opportunity to model their preferences within the current session [42][46] and across multiple sessions [34][38][45]. However, these models focus on term or topic level interests rather than degrees of belief in particular outcomes. This provides a limited view on factors that could affect preferences, and considering beliefs and biases could improve personalization. Research on result diversity (e.g., [25][40][48]) is also relevant but typically focuses on aspects such as topical variance, rather than results skewed toward one perspective or biased search behaviors.

The research presented in this paper extends previous work in a number of ways. First, we are the first, to our knowledge, to examine biased beliefs and skewed search engine result lists in IR. Second, through a detailed study of yes-no questions (involving a survey, human labeling, and large-scale log analysis – the latter two focused on the important domain of health search) we empirically demonstrate and quantify the effect of biases in results and search behaviors. Finally, we establish the effects of biases on search outcomes, specifically the accuracy of answers that people find.

3. BELIEFS DURING SEARCH

We begin by describing the retrospective survey that we performed to better understand the role of biased beliefs in Web search. This provided us with insight into people's beliefs before searching as well as afterwards, their perceptions of the process, and their rationales for their actions. An invitation to complete an online survey was distributed via email to a sample of employees within Microsoft Corp. The sample comprised employees in a range of technical and non-technical roles. We were particularly interested in those who had recently pursued yes-no answers using search engines. Since there are only two opposing outcomes, such questions provide a means of quickly and easily measuring degrees of belief. A total of 198 respondents (23.1% of all respondents) reported issuing such a question to any Web search engine during two weeks immediately prior to survey distribution. We asked participants to recall that particular searching episode, and provide the yes-no

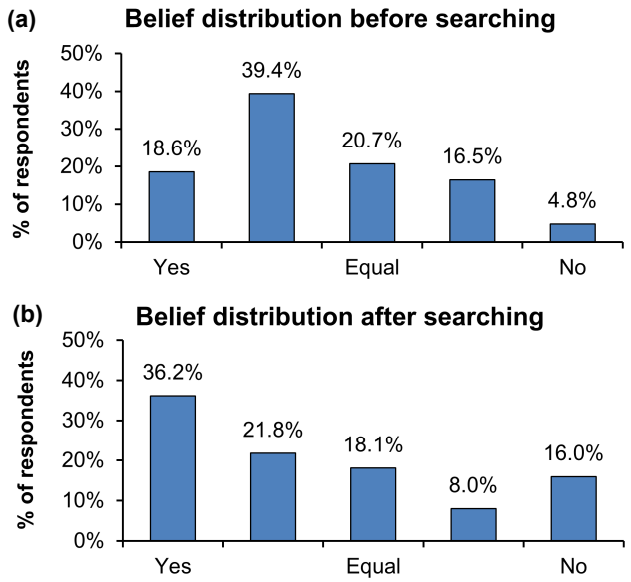

Figure 1. Distribution of (a) prior and (b) posterior reported recalled beliefs about the answer (ranging from Yes to No).

query issued, as well as the motivation behind the search. Question queries reported by respondents included "*Does chocolate contain caffeine?*" and "*Are shingles contagious?*". We also asked respondents questions about their experience during their recalled search.

We believe that there is value in using retrospective analysis to explore search beliefs, even though our method depends on participant recollections and they already knew the outcome of the recalled search episode. Alternative methods such as in-situ judging are intrusive and may draw undue attention to beliefs at query time, asking respondents to create yes-no questions at survey time lacks realism, and third-party judges issuing these questions as queries could not accurately consider searchers' true beliefs or motivations.

3.1 Belief Dynamics

The focus of our study was on how participant beliefs changed as a result of performing the search that they recalled. To do this, we asked participants to quantify their level of belief in each of the outcomes (*yes* and *no*) before and after the search.

Before searching: Given that they recalled a recent yes-no question searching episode, survey respondents were then asked to "*Rate your relative prior belief about the likelihood of each outcome before you used the search engine*" on a nine-point scale, where extreme values corresponded to the response options *yes* and *no*, and the mid-point corresponded to an equal belief in *yes* and *no*. This allowed us to obtain a precise distribution of beliefs across the range of response options. Given small counts in some cells, and to more easily identify trends, we created five response groups: *yes*, *lean yes*, *equal*, *lean no*, and *no*. The *lean yes* and *lean no* groups comprise the three ratings between *equal* and *yes* and *no* respectively. Figure 1a has the distribution of reported recalled beliefs.

Figure 1a provides evidence of a positive skew in respondents' beliefs before they search. Specifically, 58% of respondents leaned toward *yes* and only 21% leaned toward *no* (the remaining 21% reported an equal belief in both outcomes). This corroborates prior research on confirmation discussed earlier in the related work section (e.g., [26][41]). In addition, analysis of the explanations offered for leaning *yes* or *no* revealed that 47% of respondents explicitly cited *confirmation* of their beliefs as the primary search activity they were performing (e.g., "I was verifying", "I wanted to check").

Table 1. Fraction of respondents reporting different beliefs following search (columns) given the belief before (rows).

Belief before search	*Belief after search*				
	Yes		**Equal**		**No**
Yes	77.1%	8.6%	8.6%	2.9%	2.9%
	35.1% ❶	40.5%	9.5%	2.7% ❷	12.2%
Equal	23.1%	12.8%	43.6%	15.4%	5.1%
	16.1%	9.7%	19.4%	19.4% ❶	35.5%
No	11.1%	–	11.1%	–	77.8%

After searching: We also wanted to understand the impact of search on respondents' beliefs. To do so we asked respondents to "*Rate your relative posterior belief about the likelihood of each outcome once you finished searching*" The distribution of responses appears in Figure 1b. This distribution shows less uncertainty, with more responses at *yes* and *no*. One possible explanation for this is that people were more informed after they examined results. There is more of a split between *yes* and *no*, with less than half of respondents still exhibiting some uncertainty (i.e., 48% believed *lean yes*, *equal*, or *lean no*) compared to more than three quarters of respondents (77%) before searching. The fraction of searchers who believed *yes* after the search is more than double that of any other outcome, suggesting that respondents mostly shifted their prior beliefs from *lean yes* to *yes*. However, this analysis is insufficient to understand the belief dynamics in individual respondents.

Table 1 shows the fraction of users transitioning to a particular belief after they search (columns) conditioned on their beliefs with respect to *yes* and *no* before they searched (rows). For example, 77% of searchers who strongly believed *yes* beforehand, still believe this afterwards, with a similar value (78%) for *no*. These findings suggest that if searchers are certain about their beliefs initially then they are likely to remain unchanged following the search. Research on anchoring-and-adjustment has shown that people typically perform little revision to their beliefs, especially if those beliefs are strongly held [21]. To our knowledge, our work is the first demonstration of this heuristic being applied in a search setting.

There are other noteworthy findings from Table 1:

1. Respondents leaning *yes* or *no* before searching either retain that belief, or become more certain (i.e., ① in Table 1).
2. Respondents with a belief in *yes* or *no* (definite or lean) at the outset are unlikely to dramatically change their opinion to the opposing perspective after the search (e.g., ② in Table 1).
3. Respondents who are uncertain at the outset (*equal*) are likely to either remain uncertain (44%) or be more likely to lean positive following the search, e.g., 36% of respondents (23.1% + 12.8%) who believed equal beforehand believed *yes* afterward, versus 21% who believed *no* (15.4% + 5.1%).

There are a few possible explanations for these findings: (i) searchers are drawn to information supporting or confirming their prior beliefs (as suggested in survey remarks) and are therefore unlikely to change their opinion as they are not exposed to contradictory evidence, (ii) search engines rank results with *yes* higher in the list leading searchers to be more likely to view those results, or (iii) the prior distribution of correct answers to the yes-no questions submitted to search engines is skewed positive. We answer the first question using survey responses and later log analysis. The others require an assessment of the content of the SERPs and landing pages, and a correct answer to the questions posed. We do not address these directly in the survey, since respondents may not be able to make the assessments objectively. However, later in the paper we describe how we obtained direct labels for SERPs and results and answers for a subset of yes-no question queries.

Table 2. Reported motivations for considering other answers once candidate answer was found. Multiple reasons permitted.

Reason	*Percentage of responses*
Confirmation of the current best answer from another source	83.9%
Information that contradicted the current best answer (e.g., to assess its validity)	43.4%
Direct comparisons of the two outcomes (*yes* and *no*) in the same Web page	16.9%

3.2 Answer Perceptions and Follow-on Search

Answers found using search engines can affect action in the world. It is therefore important to understand whether people report finding answers and how confident they are in the accuracy of those answers. Overall, 85% of respondents reported finding an answer to their yes-no question by searching, with 92% of those who found an answer reporting that they were *confident* or *extremely confident* in answer accuracy. These high percentages are encouraging if the search engine is correct, but as we show this is not always the case.

Since searchers frequently examine multiple results [11], we were also interested in the motivation behind additional seeking once an initial answer was found. We asked respondents: "*If you found an answer early in your search, did you still consider multiple results before settling on your final answer?*" In total, 49% of our respondents reported viewing multiple answers. We then asked them to explain the motivation behind that additional searching (Table 2). The findings in the table demonstrate that once a searcher finds an initial answer their follow-up searching is likely to be related to seeking confirmatory information. Overall, 84% of respondents reported that confirmation of the initial answer was their motivation, compared to the 43% and 17% who provided reasons linked to testing the robustness of the initial answer with contradictory information.

3.3 Summary

The survey was part of an initial exploration of belief dynamics and their impact on search behavior. There are two main takeaways from the findings that are relevant to our focus in this paper:

1. Respondents reported confirmatory search behaviors by sticking to an answer if they strongly believed *yes* or *no*, transitioning to a stronger belief if they leant toward *yes* or *no* initially, respondents favored information supporting their belief, and sought confirmation once an answer was found.
2. If people are unsure (i.e., have equal belief in *yes* and *no*) they are almost twice as likely to move toward a positive answer rather than negative following searching.

Search engines may be supporting these biased processes and we need to explore this further as part of our study. While the survey findings provide motivation to study search-related biases, to study them in detail we need ground truth answers for the questions that people pose and judgments reflecting the results surfaced by search engines. We performed an extensive follow-up study using human labeling of SERPs and results, and a log analysis of search behavior for yes-no questions. Focusing on the medical domain, given its importance and since we could gather ground truth answers, we obtained answers from physicians and judgments from crowdworkers.

4. QUESTIONS, ANSWERS, JUDGMENTS

We first describe the automatic identification of yes-no questions from logs, the process of obtaining answers to a subset of those questions from physicians, and the process of gathering judgments about answers in SERP captions and landing pages for that subset.

4.1 Searcher Questions

We automatically extracted yes-no questions from a random sample of the logs of queries issued by 2.3M users of Microsoft Bing during a two-week period from September 2012. The data includes user identifiers, timestamps, queries, result clicks, and the captions (titles, snippets, URLs) of each of the top 10 results. To remove variability from cultural and linguistic variation in search behavior, we only include log entries from searchers in the English-speaking United States locale. We also obtained the HTML content of each of the top 10 results from the engine index during the same period.

4.1.1 Extracting Questions

Yes-no questions are an interrogative construction where an answer of *yes* or *no* is required. We automatically analyze the logs searching for 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 non-yes-no questions from our data. During extraction, we normalized queries via lowercasing, whitespace trimming, and punctuation removal. We identified 3.4M yes-no question queries using this method. The yes-no question queries in our set comprised around 2% of the total query volume in our sample and covered a wide range of topics. Since we are interested in the assessment of answer accuracy, we needed to filter the yes-no questions to a topic where we could obtain reliable answers. To this end, we restricted the questions to those with a medical intent and sought answers to them from trained medical professionals (physicians). To help ensure data quality, we did the following additional filtering: (i) selected SERPs with same 10 results and same result ordering across all instances of the query in the two weeks, and; (ii) focused on query instances that were either the only query in the session or the terminal query in the session with no preceding queries with query-term overlap. Filter #2 gave us more certainty that users had terminated their search with that query. This is important in later analysis when we infer answer attainment from clicked results.

4.1.2 Labeling Questions with Medical Intent

Given the large number of questions selected as described in the previous section, we needed a way to automatically label whether questions had medical intent. To do this we used a proprietary classifier from Bing. The classifier labeled 2.5% of yes-no questions as having strong medical intent (threshold > 0.8). This aligns with prior analysis of query topics, which showed that approximately 3% of search queries are medical [45]. From this set, we randomly selected 1000 medical yes-no questions. To remove noisy queries and provide sufficient data from which to analyze search behavior, selected questions had to be issued by at least 10 users. Examples of the questions selected include "*Do food allergies make you tired?*", "*Is congestive heart failure a heart attack?*", and "*Can aspirin cause blood in urine?*". We restricted the size of the question set to 1000 questions since we wanted to obtain answer labels from medical professionals, who were time constrained.

4.2 Physician Answers

As stated earlier, to measure biases we needed ground truth answers to the yes-no questions in our set. We employed two practicing physicians as judges. Each judge reviewed the same set of 1000 medically-focused yes-no questions and provided an answer on a three-point scale: *yes*, *50/50*, *no*. The judges worked independently and there was no opportunity for discussion between them to resolve disagreements. Judges were encouraged to apply their knowledge and think of the most common scenario/circumstances that could apply when a user types such a question on the Internet. The middle

rating (50/50) was only to be used if: (i) there really was 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 an answer. Two other response options were provided: (i) *don't know*: if the judge did not know the answer, and (ii) *n/a*: if the query was not medical or was not a yes-no question (possible given the automated identification of yes-no questions from logs).

Overall, 4% of the questions were labeled by at least one of the two judges as not being a yes-no question. These questions were excluded from further analysis. Of the remaining 960 questions, the judges agreed on either *yes* or *no* as the answer to the question for 674 (70.2%) of them. Seeking a second opinion is common practice in medicine and the 30% disagreement between physicians is similar to that reported in medical literature [18]. However, if we consider the disagreement in more detail, we see that for 14.0 % of the questions (roughly half of the disagreement), judges completely disagreed on the answer (i.e., one labeled *yes* and the other chose *no*). For 15.8% of the questions, one judge was unsure (13.6%) or both were unsure (2.2%). The agreement matrix is in Table 3.

The percent of overall agreement across all of the four answer options was 72.2%. The Cohen's free-marginal kappa (κ) inter-rater agreement, considering chance agreement between raters is 0.630, signifying substantial agreement. Note that we use the free-marginal kappa because our raters were not forced to assign a certain number of cases to each category [7]. If we only focus on the questions where both judges provided a *yes* or *no* response, the percent of overall agreement rises to 83.4%, with κ=0.668, signifying even more substantial agreement. The only discernible difference between the questions where the physicians disagreed directly on *yes* and *no*, was that those with disagreement were much more likely to start with "can" (e.g., "*Can a pinched nerve in neck cause throat pain?*"). Overall, 49.3% of questions with disagreement started with "can," vs. 34.0% of questions with agreement. Since the usage of "can" denotes *possibility*, these questions may be more subjective and dependent on the knowledge and experience of the judges.

Since there is a high amount of uncertainty for the *50/50* and the *don't know* categories, and they occur infrequently (e.g., only 12 cases where the experts both believed that more information was definitely required), we focus on the cases where both judges were sufficiently confident to assign a rating of *yes* or *no*. Because the questions with yes-no disagreement might highlight contentious, subjective, or difficult questions, we focused on the 674 questions where both judges agreed that the answer was *yes* or *no*. Within this set, 55.2% of answers were *yes* and 44.8% *no*. These particular percentages are important because they provide the background probabilities, or *base rates*, of each answer across our data set.

Returning to our earlier discussion of why post-search beliefs tend positive (Section 3), this provides supporting evidence that the questions being posed are more likely to be answered affirmatively, although the difference is not as large as the 2:1 ratio between *yes* and *no* shown in Figure 1b, perhaps because the question sets differ. Using search logs, we have the ability to analyze results and behaviors in detail. If search engines or users lean toward *yes*/*no* significantly more than the 55/45 split, this offers direct evidence of bias.

4.3 Crowdsourced Judgments

So that we could assess the level of skew in SERPs and results, and also study which captions and results people selected, we required judgments on the perspectives offered by captions and results. For this task we used crowd-sourced judges from a pool provided under contract to our organization by Clickworker.com. To suit the geographic and linguistic filtering performed on the question queries, all judges resided in the United States and were fluent in English.

Table 3. Agreement matrix in the responses from the two physicians. Highlighted = yes-no agreement between judges.

		Physician 2				
		Yes	***No***	***50/50***	***Don't know***	***Total***
Physician 1	***Yes***	**38.8%**	8.2%	3.7%	0.5%	51.2%
	No	5.7%	**31.5%**	1.2%	0.2%	38.5%
	50/50	1.8%	2.0%	1.3%	0.0%	5.0%
	Don't know	1.3%	3.1%	0.2%	0.7%	5.3%
	Total	47.5%	44.8%	6.3%	1.5%	100.0%

The judges were required to read task guidelines and successfully complete a qualification test similar before they could start judging.

4.3.1 SERP Caption Judgments

The task required judges to assess whether a caption (title, snippet, and URL) presented on a SERP 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 suggest an answer somewhat 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 provide one of the following ratings about answers to the question:

1. Yes only (affirmative): Caption *only contains* content answering the yes-no question affirmatively.
2. No only (negative): Caption *only contains* content answering the yes-no question negatively.
3. Both (affirmative *and* negative): Caption contains content answering the yes-no question affirmatively *and* content answering the yes-no question answering negatively.
4. Neither: Caption contains neither affirmative nor negative answers to the yes-no question.

Figure 2 shows examples of each of the caption labels for a variety of queries. These examples are taken from the guidelines provided to crowd-sourced judges to help them understand the task.

Between three and five judges rated each of the 6,740 captions under consideration to achieve a consensus comprising at least three judges with the same rating. In total, consensus was reached for 96% of captions, with 85% of captions attaining agreement with only three judges. We only use those captions with agreement.

4.3.2 Landing Page Judgments

In addition to judging the captions that the search engine presented on the SERP, we employed a similar methodology to judge the full text of the results returned by the search engine. As mentioned earlier, the results were retrieved from the search engine's index at the time that the query logs were recorded. We once again used crowd-sourced judges from the same pool, between three and five judges (three needed to attain consensus), as with the captions. Judges labeled each page based on whether it contained an answer and the type of answer it contained. The judges provided one of the same four ratings as they provided for the captions. Judges could provide an error label if the page could not be loaded, although this was seldom used (for < 2% pages). In total, consensus was achieved for 92% of pages (excluding errors); 81% of results had agreement among three judges and did not need further judgments. This labeling task was slightly more difficult than caption labeling because the full page had to be inspected and some answers may be missed.

4.3.3 Validating Crowdsourced Judgments

We suspected that these tasks (which were recognition oriented) would not require the specialized medical training. To verify the correctness of this decision we asked our medical experts and our

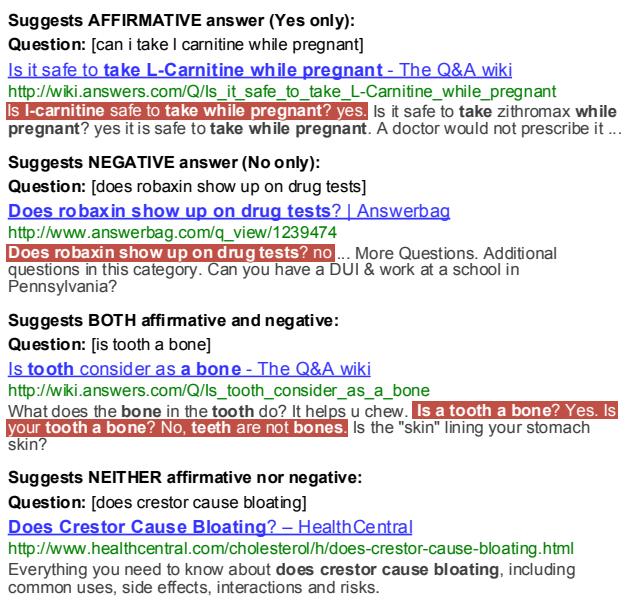

Figure 2. Examples of each of the caption ratings. The answer text is highlighted in the first three captions.

crowd-sourced judges to label the same sample of 100 captions and 100 landing pages with the labels described above, and we assessed the agreement between them. We found a high level of agreement between the physicians (both $\kappa \geq 0.886$), as well as between the two physicians and the consensus labels provided by the crowd-sourced judges (both Fleiss' multi-rater $\kappa \geq 0.853$ [13]). The findings suggest that for these tasks, the crowd-sourced judges provide similar rating quality to trained physicians (mirroring [34]). There were only 6% labeling discrepancies between captions and their associated landing pages (e.g., caption label = *both*, page label = *yes*-only). Explanations for these differences include caption generation effects or differences in judges used for caption and result labeling.

The data described in this section allows us to study the distribution of answers on SERPs and in the results themselves, as well as search interaction. Given the physician answers, we can also analyze the accuracy of the answer pages found and condition analyses of SERPs, results, and behaviors on the ground truth answers.

5. BIASED SERPS & SEARCH RESULTS

Given the data described in the previous section we can analyze the SERPs generated by the search engine (using the caption judgments) and the top-ranked results returned by the engine (using the landing-page judgments). In this analysis we focus on the presence and distribution of *yes* and *no* answers across the 674 queries in the agreement set. Recall from earlier that we focused on queries where the SERP was unchanged over the two weeks of logs, giving us exactly one SERP per query. Studying the role of captions is particularly important as they play a critical role in searchers' decision making [9][47]. We used the physician answer to study significance in two ways: (i) changes in the distributions across answer types when conditioned on the answer types (e.g., does fraction of *yes*-only top-ranked results increase when the ground truth (from expert consensus) is *yes*?), and (ii) how closely does the distribution of answer types model the truth? For (ii), if there was no bias, the distribution of *yes* and *no* would resemble the 55/45 split in our data.

5.1 Answer Presence and Result Distributions

We first wanted to determine what fraction of the results (SERP captions and landing pages) contained an answer of *yes*, *no*, *both*, and *neither* to the yes-no question posed by the user. Table 4 shows the fraction of SERPs that contain each type of answer, in terms of occurrences in captions and occurrences in the full-text of the results. The table clearly shows a bias toward *yes*. Since these are per-SERP statistics they do not reveal much about the concentration of the answers in the SERP or their relative ordering by the search engine. Table 5 presents the fraction of all captions and results for the yes-no question containing an answer. The findings show that results and captions containing *yes* are much more likely to appear in the results than those with no; around 3-4 times more likely across all captions and all SERPs. There were no significant variations in those percentages when we conditioned using the physician answer, suggesting that for these questions the search engine is insensitive to the truth (very little change in the values as we varied the physician answer). Focusing on yes-only and no-only captions and results, McNemar's chi-squared tests and *Z*-tests of proportions showed differences between yes-only and no-only in all cases (all $p < 0.01$) and differences from the base rates (i.e., 55/45) at $p < 0.01$ in all cases other than *no*-only in Table 5 ($p = 0.54$), which was similar to the *no* answer prior in our dataset (i.e., 45%).

Table 4. Percentage of SERPs with at least once instance of each type of answer in captions or results. *N*=674. SERPs can contain multiple answers so rows do not sum to one.

Source	*Yes only*	*No only*	*Both*	*Neither*
Caption	75.5%	43.0%	17.1%	85.6%
Result	74.5%	46.1%	34.5%	95.5%

Table 5. Percentage of captions/results with answer. *N*=6491.

Source	*Yes only*	*No only*	*Both*	*Neither*
Caption	28.7%	8.4%	2.7%	60.2%
Result	35.0%	12.7%	6.3%	41.0%

Table 6. Average rank of the highest-ranked result with each answer type, including standard deviation. Top rank=1.

Source	*Yes only*	*No only*	*Both*	*Neither*
Caption	2.66±1.72	2.98±2.01	4.02±2.35	2.15±2.00
Result	2.17±1.65	2.98±1.98	3.93±2.35	3.11±2.17

Table 7. Percentage of SERPs where top *yes* caption or result appears above (nearer the top of the ranking than) the top *no*.

Source	*Yes above No*	*No above Yes*
Caption	65.1%	34.9%
Result	62.4%	37.6%

5.2 Distribution of Highest-Ranked Answers

The previous section examined answer *presence* on SERPs and in results. In this section we examine the distribution of answers in captions and results as a function of *rank* position within the top 10 results. Table 6 presents the average rank position of the first answer of each type in the result list. We focus on the first answer to help ameliorate the effects from different volumes of answer types in the top 10 results. Since people inspect the results from top to bottom [20], the first occurrence of each answer in the captions is particularly important anyway. The table shows that the topmost captions and the results with *yes* are ranked above those with *no* and other answer labels. We performed one-way analyses of variance (ANOVA) for each of the two sources (captions and results), and the findings revealed significant differences between the rank of *yes* and *no* and all of the other answers for both captions and the full text of the results (both $F(1,4665) \geq 6.85$, both $p < 0.01$). The analysis also shows that that results with *yes*-only appear higher in the ranking (rank=2.15) than captions (rank=2.66) ($p < 0.001$).

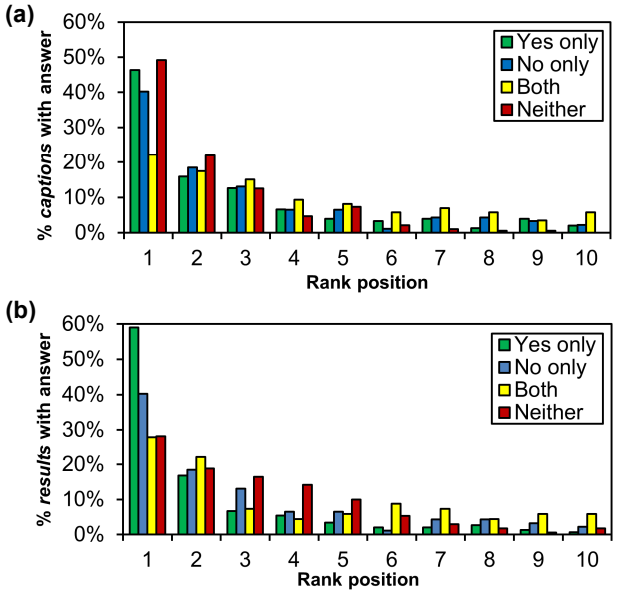

Figure 3. Distribution of *highest-ranked* answers of each type across (a) top 10 captions and (b) top 10 results.

We also computed the *distribution* of ranks for highest-ranked answers and report this in Figure 3 for answers of each type and for captions and results. The presence of an answer at ranks 2-10 is conditioned on the absence of the same type of answer at higher ranks. Figure 3a shows that the distributions of highest-ranked *yes* and *no* captions was similar across all ranks, with 40-45% occurring at the top of the list. The distribution of top-ranked answers in *results*, shown in Figure 3b, was more skewed toward the top of the ranking for *yes* than for other types, and *yes* was significantly more likely to appear in the top position than *no* (59.1% *yes* vs. 40.2% *no*, $Z = 5.00$, $p < 0.001$), once again illustrating positive rank bias.

5.3 Relative Ordering of Yes and No

Despite the evidence of bias toward *yes* answers presented in the previous section, a more direct comparison of *yes* and *no* comes from their relative ordering when both are present in the results. Our analysis also shows that around 35% of SERPs contained both a *yes* and a *no* answer. Table 7 presents the fraction of captions and results that have the top *yes* above the top *no*, and vice versa. The results show that the captions with *yes* are ranked well above those with *no* for both captions and results. The results also show that captions exhibit this bias slightly more than the results themselves (65% for captions vs. 62% for results), highlighting possible skew in the caption generation methods or rankings associated with caption clickthrough. There were no significant effects on the distribution from physician answers. However, *yes* was ranked above *no* more often than the 55% that would be expected given the base rates (both $Z \geq 2.00$, both $p \leq 0.02$), further confirming a positive bias in SERPs and results for the queries in our dataset.

5.4 Summary

Overall, we have shown that search engines are more likely to present captions/results answering a yes-no question positively (*yes*) in the top results, and they are more likely to rank results with positive answers at the top of the list, and above *no* (when both are shown on the SERP). We also showed that the distribution of answer content was insensitive to the ground truth answers provided by medical professionals; the positive bias persists irrespective of the ground truth. Finally, we showed that distributions were skewed more positive than expected from base rates. It might be argued that this bias is reasonable if searchers are satisfied. However, if search engines lead users to incorrect answers, then this requires corrective action. We examine answer accuracy at the end of the next section.

Table 8. SERP click likelihoods for different captions given variations in answer presence in SERPs/captions, and rank.

Condition(s)	*All*	*Physician Answer*	
		Yes	*No*
$SERP_Y$	80.0%	83.6%	74.4%
$SERP_N$	75.9%	80.6%	69.6%
$SERP_{BOTH}, Caption_Y$	45.6%	58.2%	41.1%
$SERP_{BOTH}, Caption_N$	14.2%	13.4%	21.8%
$Caption_Y$	41.1%	47.1%	37.7%
$Caption_N$	16.3%	12.9%	20.2%
$Caption_{Y,r=1}$	47.4%	60.0%	40.9%
$Caption_{N,r=1}$	12.6%	11.1%	17.6%

Table 9. Distribution of clicks and skips by answer. *N*=245.

Click	*Skip*	
	Yes only	*No only*
Yes only	33.3%	41.5%
No only	8.5%	16.7%

6. BIASED BEHAVIORS & OUTCOMES

We now study biases in behavior, linking SERP caption judgments with clicks, and using page judgments to study answer accuracy.

6.1 SERP Behaviors

From our search logs, we extracted all instances of the 674 yes-no queries that resulted in at least one click on a result: 496 (83%) of our yes-no question queries met this criterion. In this analysis, we focus on the *first* click that users perform and the content of the captions that they click. We do this to help reduce the effects of learning during search. We consider a number of aspects of click behavior, namely all clicks, clicks controlled for position and the distribution of yes-no results, and we also examine caption skips.

6.1.1 Result Clicks

Across all observed clicks in our set, 41.1% are on a caption labeled yes-only and 16.3% are on a caption labeled no-only (*note: the other clicks are on captions with both or neither*). Focusing solely on the clicks on *yes* and *no* that means 71.6% of clicks were on *yes*, whereas only 28.4% of clicks were on *no*. This strongly suggests that people are more likely to click on captions with positive outcomes. Table 8 provides a series of click likelihoods computed empirically from the data depicting the relationship between properties of the SERP and SERP captions, and click likelihoods (in the *All* column). The presence of *yes* or *no* in the caption is denoted as *Y* or *N* respectively, and the rank of the click is denoted r. $SERP_{BOTH}$ means that captions with yes-only and no-only answers appear on the SERP. The table also breaks down the clicks by physician answer, to help us understand the effect of the truth on click behavior as we did in previous sections. A number of conclusions can be drawn from the findings. The first is that people are more likely to click when the SERP contains *yes*, and also are more likely to click when the physician answer is *yes*, irrespective of the SERP content (top two rows of Table 8), although this difference was not significant ($p = 0.63$). Focusing on captions, people appear 2-4 times as likely to click on captions with positive content (rows 3-6), even though *yes* is only marginally more likely in our ground truth data. Considering physician answers we see that likelihoods shift as expected (e.g., clicking *no* becomes more likely when answer is *no*), but clicking *yes* is still around twice as likely (all $Z \geq 3.16$, all $p \leq 0.001$) and far exceeds what is expected given our base rates.

Table 10. Breakdown in correctness by answer definition.

Answer defn.	*All*	*Physician Answer*	
		Yes	*No*
Top result	45.0%	57.1%	22.9%
First click	50.0%	59.1%	27.9%
Last click	52.3%	66.2%	29.4%

Table 11. Distribution of changes in the answer rating for first click versus the last click for the yes-no query instance.

First click	*Last click*	
	Yes only	**No only**
Yes only	70.2%	0.0%
No only	7.2%	22.6%

6.1.2 Controlling for Rank Position

One major factor that could explain the strong preference for positive information is the tendency of the search engine to return positive search results above negative results (as shown in Section 5). Given how searchers typically examine result lists (top-to-bottom), this could lead to an *apparent* preference for positive information caused by the search engine, even if one did not exist for searchers. Since we were performing our analysis retrospectively, we could not use methods such as FairPairs [29] to counteract positional bias. To do this in our study, we focused only on the result at the top position in the list (r=1) and assume that it is always examined (a hypothesis supported by gaze tracking studies [20]). Although the distribution of *yes* and *no* answers in these captions was very similar there were still more *yes* than *no* captions at the first position (see Figure 3a for the distribution). To ensure that the distribution was equal, we randomly down-sampled the *yes* instances. This gave us the same number of *yes* and *no* captions on which to study clicks (N=357). The findings appear in the last two rows of Table 8. The trends are similar to the original analysis, and the differences are still significant (all $p < 0.001$), although the effects of the physician answer are amplified when removing position effects (e.g., click *yes* is five times as likely as click *no* when physician answer is *yes*, and still more than twice as likely when the answer is *no*).

6.1.3 Skips and Clicks

Beyond clickthrough behavior, it is also worth considering the nature of the results that users *skipped* over prior to clicking on a particular caption. This provides additional insight into searcher preferences that is not available in clicks alone. To study this in our context, we targeted SERPs with both yes-only and no-only answers in captions, and identified clicks on a caption with a *yes* or *no* answer where the user had skipped over another caption prior to clicking (e.g., skip *yes*, click *no*). Table 9 shows the percentage of all skip-click pairs in each combination. A McNemar's chi-squared test shows significant differences between the outcomes ($\chi^2(1) = 16.99$, $p < 0.001$). The table shows that it is common to skip over a caption and click on another caption with the same answer. However, what is most interesting is that on 42% of observed skip events, people skip over a caption with *no* to click on a caption with *yes*. This happens more frequently (five times) than it occurs in the opposite direction (skip *yes*, click *no*) and offers more evidence that people are drawn to information related to positive (*yes*) answers.

6.2 Answer Accuracy

Moving beyond search behavior, we now focus on whether searchers found the correct answer to their yes-no question and try to better understand the role of the engine in getting them to that answer.

6.2.1 Identifying Answers

To study accuracy in our logs, we first had to identify the particular answer that searchers found. Since we did not have direct judgments about if and where an answer was located, we inferred that from results and behavior. We devised three answer definitions:

1. Top result provided by the search engine. This allows us to study how well the search engine performs in finding answers. Note that even though the search engine may not be trying to answer the question directly, positional biases mean that searchers are more likely to select the top result [20].
2. First satisfied result click for the query instance. A satisfied click has a subsequent dwell time of 30 or more seconds, or the last click in the session [5]. This allows us to study the accuracy of the answers that searchers initially find.
3. Last satisfied result click for the query instance. This helps study the effect of user experience during the query session.

Note that since most of the impressions had only one click, then the first and the last clicks are often the same. For this analysis, we only focused on clicked pages assigned a *yes*-only or *no*-only judgment.

6.2.2 Answer Correctness

Given that we know the physician answer to each of the questions in the set (which we regard as our ground truth), and we know the answer rating for the page that searchers clicked on, or the top-ranked search result provided by the engine, we can calculate the *correctness* of the answer that the user found. Table 10 presents the breakdown of correctness by our three answer types. The findings summarized in the table show that if searchers trust the top-ranked result of the search engine, they will obtain the correct answer less than half of the time (45%). Analyzing results by physician answer, we can better understand its effect on answer correctness. In all cases, the answer was more likely be correct if the physician answer was *yes* (all $Z \geq 4.23$, all $p \leq 0.001$). If the physician answer is *yes*, then searchers settled on *yes* 66% of the time, perhaps because results with *yes* are more likely to be ranked higher in the list making them more likely to be chosen. Importantly, if the physician answer is *no*, then users attain the correct answer 23-29% of the time, showing that the focus is still on positive information.

In analyzing the results that searchers select rather than the top result that the search engine returns, we see that there is a small increase in correctness (45% to 50-52%) that can be attributed to searchers overriding the search engine ranking. The gains are small, perhaps because choices are limited to the results available in the top 10. Further analyzing the questions with high and low accuracy in the answer provided by the top result, we see that those questions starting with the terms "is" and "does" have the highest answer accuracy (61% and 59% respectively). However, questions starting with "can" (e.g., "*Can acid reflux cause back pain?*") had the lowest accuracy (38%). As mentioned earlier in the analysis of the expert labels, "can" denotes possibility and as such, for a question of this type base rates may need to factor in result ranking.

6.2.3 Answer Transitions

The findings in the survey suggested that there may be some interesting dynamics between the first and last clicks for a query (e.g., people reported being much more likely to search for confirmatory information than for information that challenged their hypothesis). We studied the changes in outcomes between queries where there were multiple clicks and different URLs at first and last click. Table 11 shows that people focus on a particular answer and stick with it as they review other resources ($\chi^2(1) = 70.01$, $p < 0.001$). The table also shows that no one transitioned from *yes* to *no*. These findings

align with our survey, which showed that confirmation was the primary motivation for pursuing information after the initial answer.

6.3 Summary

Our findings have shown that people are more likely to target results with positive content than expected given base rates. They also show that users are much more likely to skip over captions containing no-only answers to click on those with yes-only (42%) than vice versa (9%). We studied answer accuracy and showed that around half of the time people found the correct answer to our questions. Although the search engine is not trying to answer directly, biases limit selection options and previous studies have shown that people can interpret result order as a likelihood ordering [44]. In addition, we also see strong evidence to support our survey findings that people seek confirmatory information and rarely change answer focus.

7. DISCUSSION AND IMPLICATIONS

We have shown evidence of biases in how people seek information and in search engine rankings toward particular outcomes. Our mixed methods approach used a survey, log analysis, and analysis of labeled data from crowdworkers and physicians. The survey provided insight into people's belief dynamics during search, and the role of search engines and their search behavior in finding information that supports those beliefs. As we have shown, biases in the results provided by the search engine may lead users to incorrect answers; on average people were likely to find the wrong answer half of the time. However, there were also marked differences in accuracy related to the nature of the queries, with questions denoting answer *possibility* resulting in lower answer accuracy.

There are limitations of the research that we should acknowledge. First, we focused on a small set of carefully-selected queries of a particular question type. Such focus was necessary given the scale of the human labeling effort (over 2000 person-hours of judgment time for the crowd-sourced caption/content judging alone) and to simplify our analysis. Different types of questions, such as those with a subjective or exploratory focus, are also worth examining. It is also worth considering situations where biases in beliefs could be significant (e.g., diagnostic scenarios, controversial topics). Second, although there are similarities between the survey and our later analysis, the survey did not specifically focus on health. Finally, we focused on the questions with physician agreement. The questions with diverging opinions may represent challenging or difficult topics, or cases where more information is needed prior to answering.

The research has implications for improving the design of search systems and raises important points for discussion on the role of search engines. Overall, a better understanding of how search engines rank answer pages is needed. Our findings show that the results that search engines offer are heavily skewed toward particular answers (usually positive), irrespective of the truth. As suggested earlier, this may be a consequence of search engines learning from biased, aggregated user behavior [1]. It may also relate to how users frame their queries. Studies have shown that people are more likely to frame questions positively when testing hypotheses [41]. Indeed, almost all of the yes-no queries in our dataset were framed in this way (e.g., "*Can acid reflux cause back pain?*" and not "*Is acid reflux unrelated to back pain?*"). As a result, ranking methods such as vector-space similarity [30] may consider a result with a *yes*-oriented answer (e.g., "Acid reflux can cause back pain") to be more similar to the query than a *no*-oriented answer (e.g., "Acid reflux cannot cause back pain"). Although text similarity is only one component of sophisticated search engine ranking algorithms, the higher query similarity with confirmatory content may be evident in other sources such as anchor text and document titles, and needs further investigation.

Although search results may not reflect the truth, they may reflect the dominant opinions of page authors or searchers, in which case on average people may well be satisfied with the answers provided, even if they are strictly inaccurate. Research on the "filter bubble" [27] suggests that search engines have a duty of care and that personalization filters out information that disagrees with user viewpoints. In this work, we show that related effects are observed beyond personalization; search engines exhibit positive biases in result ordering irrespective of the truth. Rank ordering is important given that people may interpret ranked lists of results as a ranking of likelihoods [44]. We also show that users are more likely to engage with positively-leaning content, important because it confirms their beliefs and biases (as our survey shows). The tradeoff between helping users validate beliefs and providing accurate information, and the incorporation of likelihood information should be explored at a query, domain, user, or cohort level. Other effects such as source reliability, the ease with which people can confirm versus refute their beliefs (conclusively falsifying hypotheses can be challenging [28]), and the distribution of positive and negative answers in online content, also need to be considered. Although the general online availability of *yes* and *no* content needs to be studied, we did show in Table 7 that even when both answers are available in the top 10 results, *yes* is ranked above *no* much more frequently.

Exploring accuracy depends critically on the availability of reliable truth data. Technology developed for Web-scale question answering (e.g., [12]) may be applied in the ranking of search results to help search engines consider base rates, especially for particular queries or query classes (e.g., as we demonstrate in our analysis with yes-no questions beginning with "can"). Search engines using learning-to-rank algorithms could consider the nature of the terminology in captions associated with clicks and downweight clicks that appear to be driven by known biases, e.g., those associated with health anxiety. Focused crawling strategies, proposed in the medical domain [36], could improve the reliability of the content in the engine index, since that may affect answer correctness. Long-term models to better represent beliefs, preferences, and perspectives of users beyond topical or term-level interests may also be useful. Tools and other search support could also be provided to introduce diversity into the results, highlight more trustworthy content [32], and prominently display base rates (e.g., in answers on SERPs).

8. CONCLUSIONS

Biases affect judgment and decision making. Although these biases can impact search behavior and result ranking, they have largely been ignored in IR. We have described a detailed study of biases in search, in the context of yes-no questions in the medical domain. We employed a retrospective survey, log analysis, content labeling by human judges, and answers from physicians. We showed that people seek to confirm their beliefs with their searches and that search engines provide positively-skewed search results, irrespective of the truth. We also showed that people are more likely to select positive (*yes*) information on SERPs, even when we control for rank, and are likely to skip negative results to reach positive ones (when the opposite is not true). Perhaps the most concerning insight from our analysis is that the combination of system and searcher biases lead people to settle on incorrect answers around half of the time (and that this inaccuracy is amplified when the physician answer, used as our ground truth, is *no*). The findings highlight a tradeoff between bias and accuracy that search engines need to consider. In future work we will study biases in domains beyond health search, and on different question types. We will also investigate bias-sensitive ranking methods to tackle challenges such as when accuracy should factor in ranking, and how base rates should be accurately determined, represented, and used by search engines.

ACKNOWLEDGMENTS

The author is grateful to Peter Bailey, Susan Dumais, Eric Horvitz, and Resa Roth for feedback and discussions around related ideas.

REFERENCES

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

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

[3] Baron, J. (2007). *Thinking and Deciding*. Cambridge Press.

[4] Belkin, N.J., Oddy, R.N., and Brooks, H.M. (1982). ASK for information retrieval: Part I - background and theory. *J. Documentation*, 38(2): 61−71.

[5] Bennett, P.N. et al. (2012). Modeling the impact of short- and long-term behavior on search personalization. *Proc. SIGIR*, 185−194.

[6] Bilenko, M. and White, R.W. (2008). Mining the search trails of the surfing crowds: identifying relevant websites from user activity. *Proc. WWW*, 51–60.

[7] Brennan, R.L. and Prediger, D.J. (1981). Coefficient Kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, (41): 687–699.

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

[9] Clarke, C., Agichtein, E., Dumais, S., and White R.W. (2007). The influence of caption features on clickthrough patterns in Web search. *Proc. SIGIR*, 135–142.

[10] Craswell, N., Zoeter, O., Taylor, M., and Ramsey, B. (2008). An experimental comparison of click position-bias models. *Proc. WSDM*, 87–94.

[11] Dou, Z., Song, R., and Wen, J.R. (2007). A large-scale evaluation and analysis of personalization search strategy. *Proc. WWW*, 581–590.

[12] Dumais, S. et al. (2002). Web question answering: Is more always better? *Proc. SIGIR*, 291–298.

[13] Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5): 378–382.

[14] Fortunato, S., Flammini, A., Menczer, F., and Vespignani, A. (2006). Topical interests and the mitigation of search engine bias. *PNAS*, 103(34): 12684–12689.

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

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

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

[18] Inlander, C.B. (1993). *Good operations, Bad operations: The People's Medical Society's Guide to Surgery*. Viking Adult.

[19] Joachims, T. (2002). Optimizing search engines using clickthrough data. *Proc. SIGKDD*, 132–142.

[20] Joachims, T. et al. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM TOIS*, 25(2).

[21] Kahneman, D. and Tversky, A. (1974). Judgment under uncertainty: heuristics and biases. *Science*, 185(4157): 1214–1231.

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

[23] Kuhlthau, C. (1991). Inside the search process: Information seeking from the user's perspective. *JASIST*, 42(5): 361–371.

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

[25] Mowshowitz, A. and Kawaguchi, A. (2002). Bias on the Web. *CACM*, 45(9): 56–60.

[26] Nickerson, R.S. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Rev. Gen. Psych.*, 2(2): 175–220.

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

[28] Popper, K. (1959). *The Logic of Scientific Discovery*. Basic Books.

[29] Radlinski, F. and Joachims, T. (2006). Minimally invasive randomization for collecting unbiased preferences from clickthrough logs. *Proc. AAAI*.

[30] Salton, G., Wong, A., and Yang, C.S. (1975). A vector space model for automatic indexing. *CACM*, 18(11): 613–620.

[31] Saracevic, T. (1997). The stratified model of information retrieval interaction: Extensions and applications. *Proc. ASIS*, 34: 313–327.

[32] Schwarz, J. and Morris, M.R. (2011). Augmenting Web pages and search results to help people find trustworthy information online. *Proc. SIGCHI*, 1245–1254.

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

[34] Snow, R., O'Connor, B., Jurafsky, D., and Ng, A.Y. (2008). Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. *Proc. EMNLP*, 254–263.

[35] Sontag, D. et al. (2012). Probabilistic models for personalizing web search. *Proc. WSDM*, 433–442.

[36] Tang, T.T., Hawking, D., Craswell, N., and Griffiths, K. (2005). Focused crawling for both topical relevance and quality of medical information. *Proc. CIKM*, 147–154.

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

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

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

[40] Vaughn, L. and Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes. *IP&M*, 40(4): 693–707.

[41] Wason, P.C. (1960). On the failure to eliminate hypotheses in a conceptual task, *Q. J. of Exp. Psychology*, 12: 129–140.

[42] White, R.W., Bennett, P.N., and Dumais, S.T. (2010). Predicting short-term interests using activity-based search context. *Proc. CIKM*, 1009–1018.

[43] White, R.W. and Drucker, S.M. (2007). Investigating behavioral variability in Web search. *Proc. WWW*, 21–30.

[44] White, R.W. and Horvitz, E. (2009). Cyberchondria: Studies of the escalation of medical concerns in web search. *ACM TOIS*, 27(4): 23.

[45] White, R.W. and Horvitz, E. (2012). Studies on the onset and persistence of medical concerns in search logs. *Proc. SIGIR*, 265–274.

[46] Xiang, B. et al. (2010). Context-aware ranking in web search. *Proc. SIGIR*, 451–458.

[47] Yue, Y., Patel, R., and Roehrig, H. (2010). Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. *Proc. WWW*, 1011–1018.

[48] Zhai, C., Cohen, W.W., and Lafferty, J. (2003). Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. *Proc. SIGIR*, 10–17.