Overcoming Bias to Learn about Controversial Topics

V.G.Vinod Vydiswaran

School of Information, University of Michigan, Ann Arbor,

105 S. State St, Ann Arbor, MI 48109

vgvinodv@umich.edu (Corresponding author)

ChengXiang Zhai, Dan Roth

Department of Computer Science, University of Illinois, Urbana-Champaign,

201 N. Goodwin Avenue, MC-258, Urbana, IL 61801

czhai@illinois.edu, danr@illinois.edu

Peter Pirolli

Palo Alto Research Center

3333 Coyote Hill Road, Palo Alto, CA 94304

Peter.Pirolli@parc.com

Abstract

Deciding whether a claim is true or false often requires a deeper understanding of the evidence supporting and contradicting the claim. However, when presented with many evidence documents, users do not necessarily read and trust them uniformly. Psychologists and other researchers have shown that users tend to follow and agree with articles and sources that hold viewpoints similar to their own, a phenomenon known as confirmation bias. This suggests that when learning about a controversial topic, human biases and viewpoints about the topic may affect what is considered “trustworthy” or credible. It is an interesting challenge to build systems that can help users overcome this bias and help them decide the truthfulness of claims.

In this paper, we study various factors that enable humans to acquire additional information about controversial claims in an unbiased fashion. Specifically, we designed a user study to understand how presenting evidence with contrasting viewpoints and source expertise ratings affect how users learn from the evidence documents. We find that users do not seek contrasting viewpoints by themselves, but explicitly presenting contrasting evidence helps them get a well-rounded understanding of the topic. Furthermore, explicit knowledge of the credibility of the sources and the context in which the source provides the evidence document not only affects what users read, but also whether they perceive the document to be credible.

Introduction

The World Wide Web has become one of the primary sources of information in a variety of domains. Online news portals have gained popularity steadily over the last decade, and traditional print media is losing ground (Pew Research Center for the People and the Press, 2010). Patients and caregivers search online for health information and information about particular diseases (Kaiser Family Foundation, 2009), share their medical history (Kaiser Family Foundation, 2004), and learn about treatment options through web portals and health forums (Taylor, 2011). Students rely on online resources to complete assignments in history, literature, and other subjects. For instance, according to surveys conducted on parents of children who use computers at home, 55% of parents responded that their children spent most of the time on the home computer doing research for school, writing school work, or using education software (PSRA/Newsweek/Kaplan Education Center, 1998; PSRA/Newsweek, 2000). In all these tasks, naïve information seekers assume that the information available online is accurate, trustworthy, and unbiased. However, the Web is a hodgepodge of well-curated, edited content and freelance, unmoderated content. With more and more data and content residing in unstructured and semi-structured text format, there is a strong need to understand what is being said, and whether it can be trusted.

Typically, well-structured, formatted, and edited content is considered more trustworthy and credible (Rieh & Belkin, 1998; Wathen & Burkell, 2002) in contrast to information that appears less professional.

However, history is replete with many instances where credible sources have helped spread rumors or made significant errors in stating facts, possibly due to their own biases. If online information seekers rely only on these sources, they may get a biased view. This indicates a growing challenge to present users with enough relevant information that would encourage them to form an unbiased opinion about the topics of interest. It also provides us a strong motivation to design and build systems that would help users in this task.

Cognitive biases and their effects in decision making have been extensively studied in various branches of psychology (Baron, 2000; Plous, 1993). The closest definition of bias considered in this work is that of confirmation bias. According to (Nickerson, 1998), confirmation bias connotes the seeking or interpreting of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis at hand. Consider a scenario where an Internet surfer wants to learn about a recent controversy. To give an example, consider that Alice is a mother with young kids who are about to start school. She wants to know if chocolate and flavored milk provided in schools are a healthy food choice for her children. Depending on the keywords she chooses to search the Web about the topic, she might see news articles about a recent ban on chocolate milk in certain schools, or learn about the health benefits of milk in growing children. She might find results from news media organizations reporting on the ban, or activist groups actively encouraging drinking milk, or even concerned parents posting questions and responding to them via community-driven question answering services.

However, are all these results equally helpful and informative in satisfying the user's information need? Is Alice equally likely to read these articles, and read them in the order presented? The answers to both of these questions appear to be “no”, but we wanted to understand what factors impact these decisions. We wanted to study various factors that enable humans to acquire additional information about controversial topics in an unbiased fashion. Specifically, we wanted to study the factors that influenced (i) the documents users read, (ii) the extent of learning, and (iii) the perceived credibility of a source. This understanding will help us design better systems and interfaces to help users verify or refute controversial claims. We designed a user study to answer these research questions:

1. Does explicit display of contrasting viewpoints help users understand controversial topics better?

2. Does (knowledge of) the source rating affect the credibility judgment of the source/document?
3. Does human bias affect the credibility judgment of documents?
4. Do multiple documents and viewpoints help/hinder learning about a controversial topic?
5. Does prior knowledge/bias affect how humans learn about a controversial topic?

We believe this is one of the first works that studies these aspects, especially when learning about controversial topics. Traditional approaches to verifying controversial claims follow algorithmic approaches to assimilate possibly contentious evidence from multiple sources (Yin, Han, & Yu, 2008; Gallard, Abiteboul, Marian, & Senellart, 2010). We designed and conducted a user study, called BiasTrust, to understand how to present such evidence documents to enable users to learn about the controversy in an unbiased fashion and help them overcome any prior bias they might have about the topic.

In this user study, we first try to understand the user's position and beliefs about the controversy. Then, we study how presenting evidence with contrasting viewpoints and source expertise ratings affect how the users accessed the evidence documents. We find that users tend not to seek contrasting viewpoints, at least in a limited-time learning scenario, but that explicitly presenting contrasting evidence helps them get a well-rounded understanding of the topic. Furthermore, we observe that explicit knowledge of the sources' credibility or expertise, and the context in which the evidence was provided, not only affects what users read, but also how credible they perceive the documents to be. These insights help us optimize the presentation of credible evidence documents to teach biased users about controversial topics in the most effective way.

Related work

Understanding which documents people read is related to research in many fields. Psychologists have studied the phenomenon of *confirmation bias* (Plous, 1993; Nickerson, 1998; Baron, 2000) or selective exposure (Prior, 2003), which states that people tend to favor information that confirms their beliefs -- not only in deciding what to read, but also in how they interpret what they read about. Similarly, researchers in political science (such as Taber and Lodge, 2006) observed that people processed information in biased fashion, i.e. people were quick to critique opposing arguments, while uncritically accepting arguments that supported their own beliefs. We believe that one of the ways to help people overcome this bias is by developing tools and systems that present people with arguments from multiple viewpoints. We designed a user study to understand various factors that can help people in this goal.

Researchers have studied various aspects of credibility of online information within specialized domains, such as credibility assessments on the World Wide Web (Danielson, 2005), use of information technology across disciplines (Rieh and Danielson, 2007), and in academic disciplines such as communication (Metzger, Flanagin, Eyal, Lemus, & McCann, 2003). Rieh (2002) studied the problem of judgment of information quality and cognitive authority by observing people's behaviors on the Web and proposed its practical implications on the design of web systems. Fogg and his colleagues (Fogg, 2003; Fogg & Tseng, 1999; Fogg, et al., 2001) have studied various evaluation strategies to assess credibility of information over the Web. A few studies have also explored the problem of judging quality of information in user-generated, free-format venues such as Internet discussion forums and message boards based on content analysis (Savolainen, 2011, Vydiswaran, Zhai, & Roth, 2011). Although assessing credibility of the evidence collected from multiple sources over the Web is relevant to the research presented here, in this paper, we focus on studying how to better design interfaces to present credible information and the impact of presentation styles on readability and learning, especially about controversial topics.

Previous research has also looked at aggregating information from multiple sources to answer specific questions. Wu and Marian (2007) studied how to collect information from multiple sources over the Web, specifically to find answers using corroborative evidence from multiple sources. Gallard, Abiteboul, Marian, and Senellart (2010) looked at how to collect information from contrasting viewpoints. The work presented in this paper looks at the next step of how to present these documents with contrasting viewpoints to users, while also providing information about credibility of the sources, to enable users to learn about the topic efficiently.

Researchers have also looked at factors that influence what information users access and how they process it. This includes work on building tools to increase the transparency and credibility of Wikipedia articles (Pirolli, Wollny, & Suh, 2009; Suh, Chi, Kittur, & Pendleton, 2008) that give users a clearer sense about what information is credible and what is not. Ugander, Backstrom, Marlow, and Kleinberg (2012) studied the issue of how to convince people and looked at the influence of one's social network on their actions. Extending their argument to the information domain, this research suggests that users need to be exposed to multiple viewpoints to help them make their decision. Leiserowitz, Maibach, Roser-Renouf, & Hmielowski (2011) observe that those who reject the scientific evidence for climate change are, in fact, also those who believe that they are best informed about the subject. This implies that those who are ignorant about a topic are more accepting of trusted information as compared to those who are misinformed and hold on to false beliefs. On the other hand, Pariser (2011) investigated the notion of the *filter bubble*, where search engines personalize web search results to show users what they like to see and read, thereby showing them only information that agrees with their viewpoints. This not only supports confirmation bias, but also excludes contradictory viewpoints. Further, Lewandowsky, Ecker, Seifert, Schwarz, & Cook (2012) study how misinformation gets disseminated, why efforts to retract misinformation fail, and how corrections should be designed to maximize impact. Our study tries to understand how to overcome these shortcomings by presenting contrasting, yet credible evidence to users. We believe our user study is the first to study how expertise rating for sources, the evidence context, and contrasting viewpoints help users learn about controversial topics.

A shorter version of this work appeared in Vydiswaran, Zhai, Roth, and Pirolli (2012a), in which we discussed the decisions made in designing the user interfaces. Further, the effect on learning of presenting evidence passages in a contrastive format was discussed in Vydiswaran, Zhai, Roth, and Pirolli (2012b). This work combines the previous publications and extends it in the following directions. First, it gives a complete description of the user study and the experimental procedure, so as to give readers complete context to interpret the results. Second, we include additional analysis of the effect of expertise ratings and interface variants on agreement, informativeness, and perceived bias of passages. We find that documents with very low ratings and very high ratings are both perceived to be strongly biased. We also find that interface variants with contrastive viewpoints also help reduce previously held biases, compared to the interface variants with a single viewpoint display. Third, this work includes a detailed analysis of how participants interacted with the system and what features were found to be useful during the BiasTrust study. Finally, we summarize the feedback we received from the participants, and how it helped us design follow up studies.

BiasTrust: Designing the study

Understanding which claims to believe and why to believe those claims are important in order to make an informed decision. This is basically a learning task, where an inquisitive user tries to learn as much as possible about the claim and assimilate all evidence in support of or against the claim. This is an interesting challenge for a retrieval system, to not only retrieve documents relevant to the claim, but also present them succinctly to help users understand that information quickly. However, as we pointed out in the related work section, previous research by psychologists and others has shown that users tend to access information that supports their own viewpoints. So, it is important for an automated system to model such human biases and present trustworthy evidence to overcome this bias, where possible.

We designed a system that retrieves relevant, trustworthy documents and provides the user with an overall, unbiased perspective about the topic. In order to optimize the interface design, we conducted a user study, called BiasTrust, to investigate the factors affecting the choices humans make about what to read and the documents they judge as relevant and credible. This would help us to design and improve interfaces for an automated claim verification system that allows users to validate claims by providing contrasting evidence for and against the claim.

We focused on three major factors, viz. (i) the ability to access credible, yet contrasting viewpoints about a claim to help gauge the trustworthiness of the claim; (ii) the knowledge of source expertise and credibility, and how that affects the decision on what evidence is read; and (iii) the order in which the documents are presented, and if that affects the overall understanding of the topic. Other factors, such as summarizing documents to help faster learning and providing an overall truth value for a claim may also be relevant for designing a claim verification system. However, we chose not to study these factors, but instead chose to provide access to the relevant evidence to allow users to verify claims by themselves.

The user study was designed as a learning task, where participants are asked to learn as much as possible about a topic within a stipulated time. This setup helped participants decide, given the limited time, which sub-topics are important for them to learn about and choose which documents to read accordingly. We could then observe their actions and study how various factors helped or hindered them in this learning process.

Another decision in designing the study was to focus on controversial claims, instead of factual claims, where there is a lot of evidence supporting the claim and very few, possibly untrustworthy arguments against it. By choosing controversial topics where there is genuine evidence both supporting and opposing the claims, we could understand how preference-based factors affect the learning process.

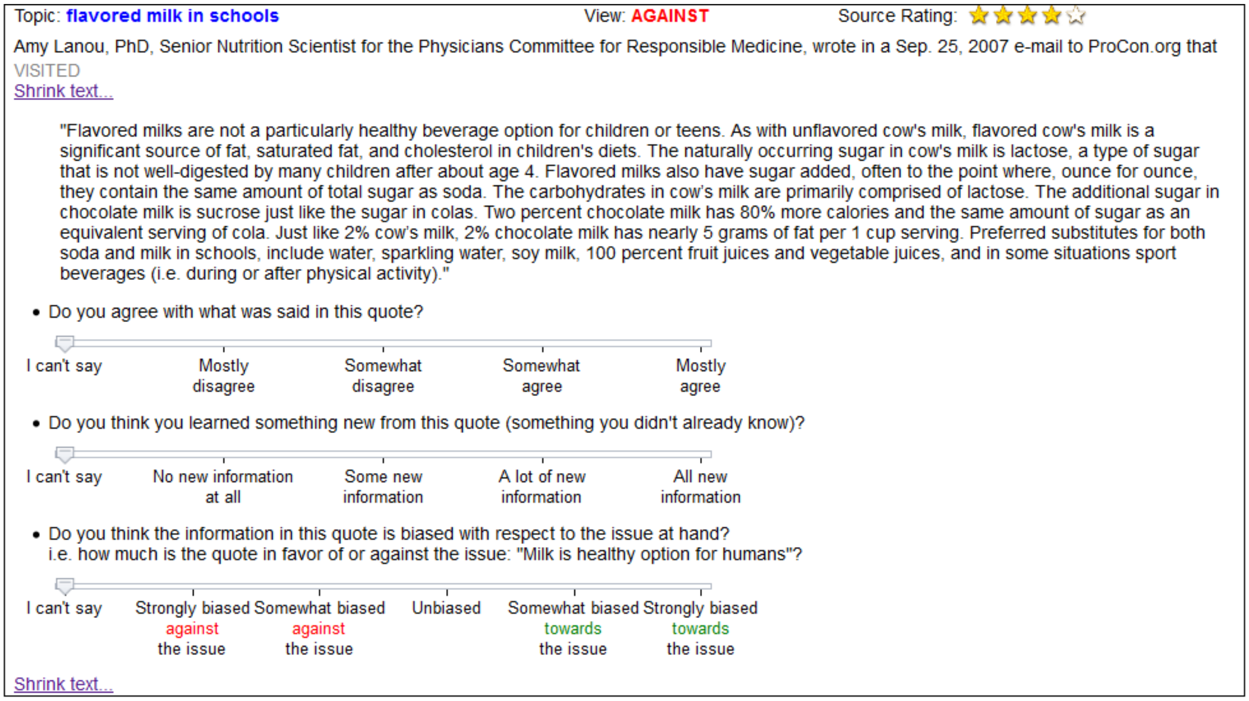
The study was conducted in four stages, viz. (i) Pre-study survey questionnaire, (ii) Study phase, (iii) Post-study questionnaire, and (iv) Feedback interview. The first three stages were conducted online, while the feedback interview was conducted face-to-face. These four stages are explained below.

Stage 1: Pre-study questionnaire

The pre-study survey questionnaire was designed to measure the participants' knowledge about the controversial topic. Specifically, participants were asked questions that helped us gauge their (lack of) knowledge and bias towards/against important issues relevant to the topic being studied. By leveraging a pre-test survey to judge the knowledge and bias about the topic, the researchers were able to study how these affect the overall selection of documents the participants want to read and the credibility judgments they make while reading new documents.

Participants answered the questions on a four-point Likert scale. For the knowledge-related questions, the scale ranged from (i) ‘insignificant’ to (iv) ‘very significant’, while for bias-related questions, this ranged from (i) ‘strongly against’ to (iv) ‘strongly in favor of’ the issue. There were also a few preference questions that were answered on a five-point Likert scale that ranged from (i) ‘strongly prefer one option’ to (iii) ‘prefer both options equally’ to (v) ‘strongly prefer the other option’. In order to capture the possible lack of knowledge about the sub-topic being discussed, participants could respond to any question with an ‘I don’t know’ answer.

This design of using knowledge and bias related questions and limiting the nature of allowed responses was intended to encourage participants to think about their position on many sub-topics related to the overall issue. For example, if the issue being discussed is whether drinking milk is healthy for humans, the questionnaire might include questions asking participants if they were aware of the issue of flavored milk being distributed in schools (knowledge question) and if they believed flavored milk is a healthy choice (bias question). In another related question, participants may be asked whether organic milk was a healthier option than conventional milk (preference question). The pre-study questionnaire also included a few demographic questions, such as age and political inclination. Further, there were a couple of task-specific questions to understand the participants' bias or preference on the issue.



Topic: **flavored milk in schools** View: **AGAINST** Source Rating: ★★★★☆

Amy Lanou, PhD, Senior Nutrition Scientist for the Physicians Committee for Responsible Medicine, wrote in a Sep. 25, 2007 e-mail to ProCon.org that

VISITED

Shrink text...

> "Flavored milks are not a particularly healthy beverage option for children or teens. As with unflavored cow's milk, flavored cow's milk is a significant source of fat, saturated fat, and cholesterol in children's diets. The naturally occurring sugar in cow's milk is lactose, a type of sugar that is not well-digested by many children after about age 4. Flavored milks also have sugar added, often to the point where, ounce for ounce, they contain the same amount of total sugar as soda. The carbohydrates in cow’s milk are primarily comprised of lactose. The additional sugar in chocolate milk is sucrose just like the sugar in colas. Two percent chocolate milk has 80% more calories and the same amount of sugar as an equivalent serving of cola. Just like 2% cow’s milk, 2% chocolate milk has nearly 5 grams of fat per 1 cup serving. Preferred substitutes for both soda and milk in schools, include water, sparkling water, soy milk, 100 percent fruit juices and vegetable juices, and in some situations sport beverages (i.e. during or after physical activity)."

- Do you agree with what was said in this quote?

- Do you think you learned something new from this quote (something you didn't already know)?

- Do you think the information in this quote is biased with respect to the issue at hand?
  i.e. how much is the quote in favor of or against the issue: "Milk is healthy option for humans"?

I can't say
Strongly biased against the issue
Somewhat biased against the issue
Unbiased
Somewhat biased towards the issue
Strongly biased towards the issue

Shrink text...

Figure 1.

Sample document view.

Stage 2: Study phase

Once the participants' responses to the pre-survey questionnaire were recorded, they were directed to one of the interface variants. The interface variants will be described in detail in the following section. In each interface variant, participants had some contextual information about the passages. For instance, participants were shown the source of the passage, the sub-topic the passage is closely related to, and whether the passage was in favor or against the sub-topic. In some variants, the expertise rating of the source was also shown to further help participants decide if they wanted to read the passage. A '*Read more…*' link was provided alongside the contextual information. Participants chose to read a passage by clicking on the '*Read more …*' link; and were then shown the passage, along with the contextual information. Figure 1 shows a sample document view.

For each passage participants read, they were asked to answer three questions about the passage, viz. (a) did they agree with what was being said in the passage; (b) did they get any new information from the passage; and (c) did they believe the information was biased with respect to the topic being discussed. These three questions allowed us to quantify the perceived importance of the passage.

Participants specified their agreement with the information given in the passage over a four-point Likert scale ranging from (i) 'mostly disagree' to (iv) 'mostly agree'. The main reason for using a four-point scale rather than a five-point one was to force the participants to decide whether they agreed with the overall passage or not.

Similarly, participants were asked to gauge how much they learned from the passage, by specifying informativeness of the information in the passage. They chose one of the options on a four-point Likert scale that ranged from (i) 'no new information at all' to (iv) 'all new information'. Participants were directed to answer the question considering the sequence of passages they read. For example, if the participant had already read three passages on the sub-topic, and then they see a fourth passage with no additional information, they were directed to respond to the informativeness question with 'no new information', even if, when considered in isolation, the passage had relevant information.

On the question of whether they believed the passage was biased with respect to the topic being discussed, participants answered using a five-point Likert scale that ranged from (i) ‘strongly biased against the topic’ to (iii) ‘unbiased’ to (v) ‘strongly biased towards the topic’. Participants were informed that the response to this question may differ from the viewpoint mentioned in the contextual information. For all three questions, participants had an option to choose ‘I can't say’ as a response, if they were not sure.

These three questions allowed us to quantify the importance participants might give to the passage they just read. By first recording the choice of the passages participants read, and then their opinion about the quality of the passage based on agreement, informativeness, and perceived bias; we could monitor how the passages helped them learn about the topic. After answering the three questions, participants could choose to read another passage. They were asked to continue reading until they believed they had read enough about the topic. Once they decided to quit the study phase, they were taken to the third and final stage of the online study.

Stage 3: Post-study questionnaire

After participants spent time learning about the topic in the study phase, they were asked to respond to a series of questions about what they learned. They were asked to specify what concepts they now believed supported the topic being studied and what concepts opposed it. They were also asked to answer the topic-specific questions that were posed during the pre-study questionnaire.

Participants responded based on the topics they read about and how important and relevant they felt the sub-topics were, after the study. Participants then wrote a short summary essay on what they learned. They were asked to spend about five minutes on the essay. They were also asked to provide feedback on what interface features helped them in their task, and suggest additional features that might help them.

Stage 4: Feedback interview

The final stage of the study was a face-to-face meeting and debriefing session. Participants were debriefed about the study and were informed about the factors that were being studied. This also provided an additional avenue for participants to provide feedback about the system and suggest changes to improve the study.

UI#	UI Variant name	# of passages	Contrast view	Topics sorted	Source rating	
					Show	Scheme
1a	**Single-Single-Bimodal-Unsort**	1	No	No	Yes	Bimodal
1b	**Single-Single-Uniform-Unsort**	1	No	No	Yes	Uniform
2a	**Single-Contrast-Bimodal-Unsort**	2	Yes	No	Yes	Bimodal
2b	**Single-Contrast-Uniform-Unsort**	2	Yes	No	Yes	Uniform
3	**Multiple-Contrast-Bimodal-Unsort**	10	Yes	No	Yes	Bimodal
4a	**Multiple-Contrast-Bimodal-Sort**	10	Yes	Yes	Yes	Bimodal
4b	**Multiple-Contrast-Uniform-Sort**	10	Yes	Yes	Yes	Uniform
5	**Multiple-Contrast-None-Sort**	10	Yes	Yes	No	--

Table 1.

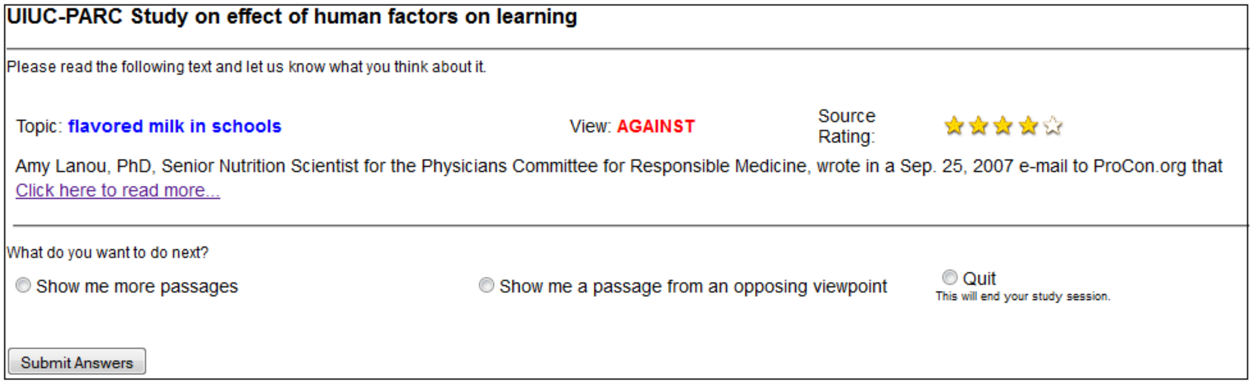
Parameter configuration for all interface variants. The variants are named based on whether **Single**/**Multiple** documents are shown, displaying **Single**/**Contrast** viewpoints, with **Bimodal**/**Uniform** rating scheme, and whether documents are **Sort**ed/**Unsort**ed based on their topics.

User interface variants

We now describe the user interface variants that we experimented with in the study. As stated earlier in the Introduction section, the main purpose of the study is to understand the factors that affect which documents are read while learning about a new topic. We designed interfaces to study the following factors:

1. Explicit display of contrasting evidence
2. Single document per page vs. multiple documents
3. Source expertise rating
4. Presentation order of documents
5. Human biases

The variants have been summarized in Table 1.



UIUC-PARC Study on effect of human factors on learning

Please read the following text and let us know what you think about it.

Topic: **flavored milk in schools** View: **AGAINST** Source Rating: ★★★★☆

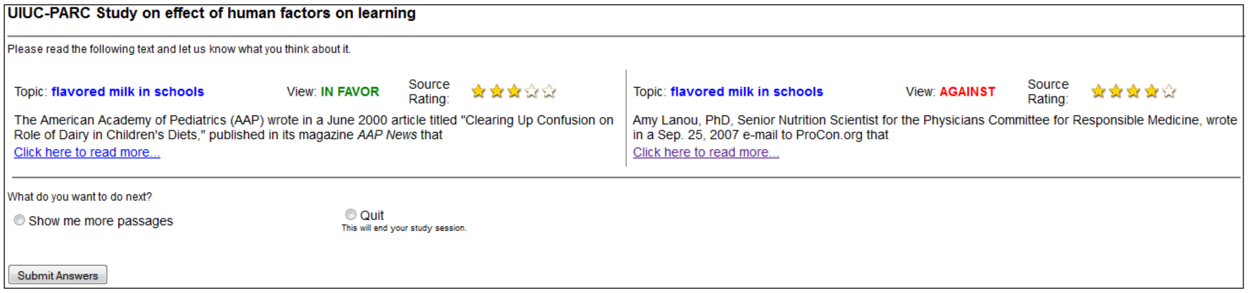
Amy Lanou, PhD, Senior Nutrition Scientist for the Physicians Committee for Responsible Medicine, wrote in a Sep. 25, 2007 e-mail to ProCon.org that
Click here to read more...

What do you want to do next?

○ Show me more passages ○ Show me a passage from an opposing viewpoint ○ Quit
This will end your study session.

Figure 2.

Single document view with option to look at contrasting document (UI variants **Single-Single-all-all**, UI# {1a, 1b}).



UIUC-PARC Study on effect of human factors on learning

Please read the following text and let us know what you think about it.

Topic: **flavored milk in schools** View: **IN FAVOR** Source Rating: ★★★☆☆

The American Academy of Pediatrics (AAP) wrote in a June 2000 article titled "Clearing Up Confusion on Role of Dairy in Children's Diets," published in its magazine *AAP News* that

Click here to read more...

Topic: **flavored milk in schools** View: **AGAINST** Source Rating: ★★★★☆

Amy Lanou, PhD, Senior Nutrition Scientist for the Physicians Committee for Responsible Medicine, wrote in a Sep. 25, 2007 e-mail to ProCon.org that

Click here to read more...

What do you want to do next?

○ Show me more passages

○ Quit

This will end your study session.

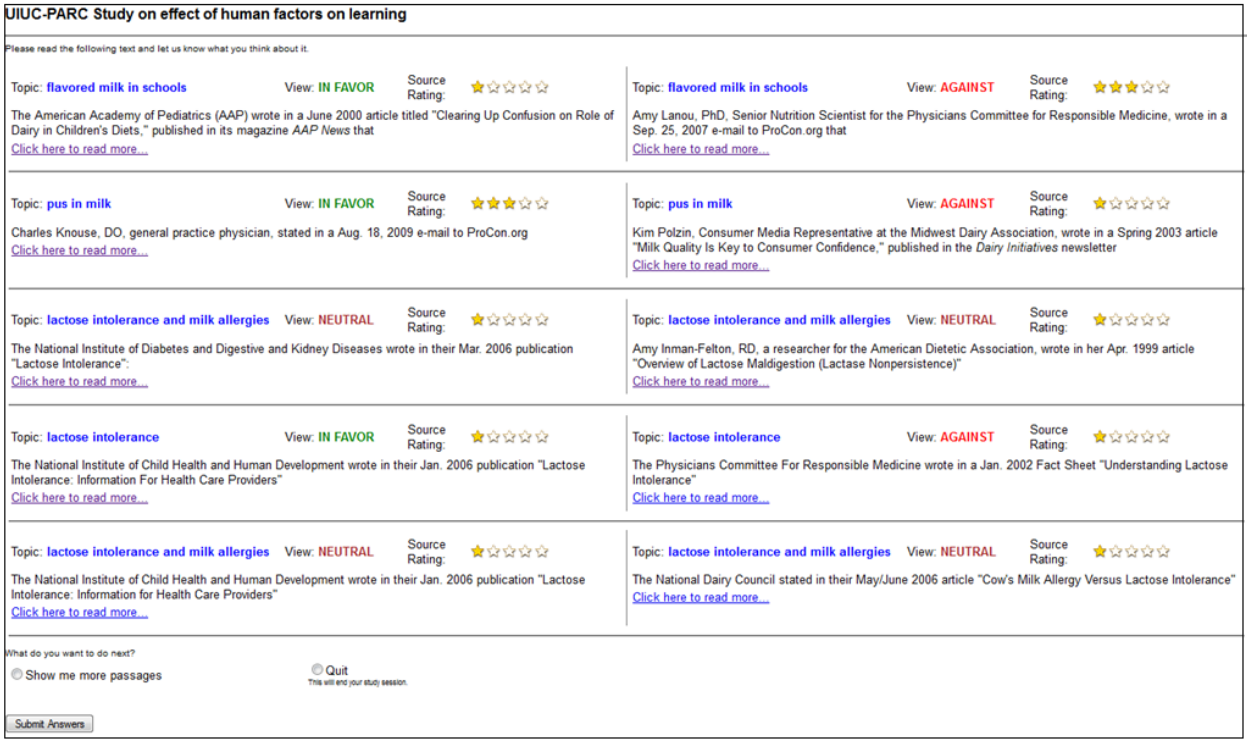
Figure 3.

Single document view that shows contrasting document by default (UI variants **Single-Contrast-all-all**, UI# {2a, 2b}).

Explicit display of contrasting evidence

One of the major factors in learning about a controversial topic in an unbiased fashion, we believe, is the exposure to alternate viewpoints. To verify if this conjecture is true, we designed two variants of the system. In the first variant, participants were exposed to just one document at a time, in a particular order. Participants had an option to explicitly ask for the next document to be one from an opposing viewpoint. If they did not choose this option, the next relevant result would be shown. Figure 2 shows an example of such an interface. UI variants **Single-Single-all-all** (UI# {1a, 1b}, cf. Table 1) follow this setting.

In the second variant, participants were exposed to the contrasting evidence right at the start. The primary document and a document of contrasting viewpoint were shown side-by-side. Participants could still pick which document to read first, and may choose to ignore the contrasting viewpoint if they wish to. Figure 3 shows an example of such an interface. UI variants **Single-Contrast-all-all** (UI# {2a, 2b}, cf. Table 1) follow this setting.



UIUC-PARC Study on effect of human factors on learning

Please read the following text and let us know what you think about it.

Topic: flavored milk in schools View: IN FAVOR Source Rating: ★☆☆☆☆

The American Academy of Pediatrics (AAP) wrote in a June 2000 article titled "Clearing Up Confusion on Role of Dairy in Children's Diets," published in its magazine *AAP News* that

Click here to read more...

Topic: flavored milk in schools View: AGAINST Source Rating: ★★★☆☆

Amy Lanou, PhD, Senior Nutrition Scientist for the Physicians Committee for Responsible Medicine, wrote in a Sep. 25, 2007 e-mail to ProCon.org that

Click here to read more...

Topic: pus in milk View: IN FAVOR Source Rating: ★★★☆☆

Charles Knouse, DO, general practice physician, stated in a Aug. 18, 2009 e-mail to ProCon.org

Click here to read more...

Topic: pus in milk View: AGAINST Source Rating: ★☆☆☆☆

Kim Polzin, Consumer Media Representative at the Midwest Dairy Association, wrote in a Spring 2003 article "Milk Quality Is Key to Consumer Confidence," published in the *Dairy Initiatives* newsletter

Click here to read more...

Topic: lactose intolerance and milk allergies View: NEUTRAL Source Rating: ★☆☆☆☆

The National Institute of Diabetes and Digestive and Kidney Diseases wrote in their Mar. 2006 publication "Lactose Intolerance":

Click here to read more...

Topic: lactose intolerance and milk allergies View: NEUTRAL Source Rating: ★☆☆☆☆

Amy Inman-Felton, RD, a researcher for the American Dietetic Association, wrote in her Apr. 1999 article "Overview of Lactose Maldigestion (Lactase Nonpersistence)"

Click here to read more...

Topic: lactose intolerance View: IN FAVOR Source Rating: ★☆☆☆☆

The National Institute of Child Health and Human Development wrote in their Jan. 2006 publication "Lactose Intolerance: Information For Health Care Providers"

Click here to read more...

Topic: lactose intolerance View: AGAINST Source Rating: ★☆☆☆☆

The Physicians Committee For Responsible Medicine wrote in a Jan. 2002 Fact Sheet "Understanding Lactose Intolerance"

Click here to read more...

Topic: lactose intolerance and milk allergies View: NEUTRAL Source Rating: ★☆☆☆☆

The National Institute of Child Health and Human Development wrote in their Jan. 2006 publication "Lactose Intolerance: Information for Health Care Providers"

Click here to read more...

Topic: lactose intolerance and milk allergies View: NEUTRAL Source Rating: ★☆☆☆☆

The National Dairy Council stated in their May/June 2006 article "Cow's Milk Allergy Versus Lactose Intolerance"

Click here to read more...

What do you want to do next?

○ Show me more passages ○ Quit
This will end your study session.

Submit Answers

Figure 4.

Multi-document view that shows five primary documents, along with the corresponding contrasting documents (UI variants **Multiple-Contrast-all-all**, UI# {3, 4a, 4b, 5}).

Single document per page vs. multiple documents

The next factor we investigated is how the amount of information on a page affects the reading pattern and choice of documents. When fewer documents are shown, humans tend to spend more time reading documents that are shown, before moving on to the next page. We wanted to investigate if this hypothesis was correct. We designed a third variant where instead of one document, five documents and their corresponding counter-arguments were shown to the user on a single page. UI variants **Multiple-Contrast-all-all** (UI# {3, 4a, 4b, 5}, cf. Table 1) show multiple documents per page, and Figure 4 shows an example.

Source expertise rating

Another factor in deciding what to read is whether users believe the document comes from a credible or trustworthy source. To test this hypothesis, we decided to control the expertise ratings in two ways. In one UI variant, the expertise ratings were hidden, and participants did not know if the source was credible or not. This setting was followed in UI variant **Multiple-Contrast-None-Sort** (UI# 5, cf. Table 1).

The second way we controlled the expertise rating was to show two different rating schemes. In one scheme, all sources got a rating of either 1 star or 3 stars. We call this the *bimodal rating scheme*, since the ratings appear to be drawn from a bimodal distribution. In the second scheme, sources were assigned a random rating drawn from a uniform distribution while ensuring that the rating is not the same as the one under the bimodal scheme. We call this the *uniform rating scheme*. Sources were assigned this rating in a static fashion; that is all participants that were shown a particular rating scheme consistently saw the same expertise rating for the same source. The four UI variants **all-all-Bimodal-all** (UI# {1a, 2a, 3, 4a}, cf. Table 1) followed the bimodal rating scheme, while the three remaining UI variants **all-all-Uniform-all** (UI# {1b, 2b, 4b}, cf. Table 1) followed the uniform rating scheme.

Document presentation order

The final hypothesis we wanted to test was whether grouping documents together by sub-topic significantly helped the learning task and affected the selection of documents read. We focused on testing this hypothesis in the case when multiple documents are shown. We had two configuration settings: in one setting, we showed the passages in the order they were retrieved (in relevance order). This would mean that the documents appear to come in a random sequence of sub-topics. Participants do not know what sub-topic the next document would come from. This setting was followed in UI variants **all-all-all-Unsort** (UI# {1a, 1b, 2a, 2b, 3}, cf. Table 1).

In the second setting, the set of retrieved documents was sorted based on topics. All documents from one topic were shown before the next topic. The topics were ordered by relevance, and the documents about a particular topic were themselves ordered by relevance. Participants could choose which documents to read from one topic before they move on to the next one. The UI variants **all-all-all-Sort** (UI# {4a, 4b, 5}, cf. Table1) were configured to follow this setting.

Setting up the user study

Data and Study topics

We enabled the study for two controversial issues, one from the health domain and the other from politics. The topics and the primary claim ("*issue at hand"*) included in this study were as follows:

1. *Milk*: Drinking milk is a healthy choice for humans.
2. *Energy*: Alternate sources of energy are a viable alternative to fossil fuels.

We will refer to the study sessions corresponding to these two issues as the *Milk* and *Energy* tasks, respectively.

We chose these particular issues because they are fairly similar in terms of biases they may invoke. We wanted topics that are controversial, but also have scientific evidence to justify either viewpoint. Notably, these issues were different from other controversial issues that may invoke strong emotional biases that are hard to overcome. Some topics that we chose to omit were regarding abortion, right to life, and nationalistic/patriotic issues. In such highly emotional issues, it is hard to find convincing scientific evidence supporting each viewpoint; and often pre-conceived notions and stands are hard to change.

For each of the two issues we included in the study, we collected over 350 snippets of text from ProCon.org (ProCon.org, 2011), a non-partisan, non-profit public charity website. A team of researchers, staff, and volunteers affiliated with the website gathered quotes from people, organizations, and other websites relevant to the issue being discussed. They grouped the quotes based on relevant questions or sub-topics within the issue, and categorized them as *pro* (in favor of the question being asked), *con* (against the question being asked), or neither *pro* nor *con*.

The website also gave an expertise rating to each source, based on the entity type. For example, governmental reports and peer-reviewed studies got the highest, 5-star rating, experts were assigned a 3- to 4-star rating, media and academic journals got a 2-star rating, while other organization and influential persons got a 1-star rating. However, our analysis showed that for the two tasks, almost all sources belonged to classes that were assigned either a 3-star or a 1-star rating. We used these manually assigned ratings as our bimodal rating scheme.

nutrition	child growth	Cancer	economy	sales
calcium	day	risk	agricultural	dairy
products	mg	cancer	food	prices
dairy	bone	milk	disparagement	marketing
fat	igf	intake	perishable	weight
iron	years	ovarian	product	milk
sources	mass	consumption	economy	orders
blood	adolescence	calcium	statutes	price
foods	childhood	women	action	minimum
vegetables	osteoporosis	associated	aquacultural	body
bone	growth	iron	damages	loss
cells	**flavored milk**	**cloning**	**environment**	**Allergy**
cows	vitamin	cloned	raw	Milk
rbst	milk	milk	emission	lactose
pus	d	manure	livestock	intolerance
cells	flavored	safety	bacteria	homogenization
hormone	children	water	milk	symptoms
consumers	daily	dairy	methane	globules
treated	source	claims	animal	protein
rbgh	nutrients	weight	beneficial	children
bst	sugar	animals	feed	lactase
fda	essential	produced	harmful	allergy

Table 2.

Topics learned for the *Milk* task.

Retrieving relevant passages

Once all passages are collected, they are indexed using the Lemur toolkit (Lemur Project, 2001). Since these passages may be from a varied set of sub-topics, we first analyze the corpus to identify key sub-topics based on a statistical topic modeling approach. For each issue, we identified ten keywords that we believed were relevant to the issue at hand. These words were used as seeds to learn a ten-topic probabilistic topic model using the probabilistic latent semantic indexing (PLSI) (Hofmann, T., 1999) approach. PLSI is a probabilistic generative model that optimizes the likelihood of generating the data as a combination of sub-topics. It analyzes the co-occurrence of words and groups them into constituent language models and in the process, learns the key sub-topics within a corpus of documents. The model can be seeded with key concepts and the algorithm learns the most likely sub-topics. Other popular approaches to learn the topics from a collection of documents include Latent Dirichlet Allocation (Blei, Ng, Jordan, & Lafferty, 2003). Table 2 shows the sub-topics learned for the *Milk* task. The titles for the topics have been assigned manually.

Once the sub-topics are learned, key terms are extracted from each sub-topic. A weighted combination of these key terms is used as a retrieval query to extract relevant evidence passages from the corpus. In the general system, the combination weights can be set specific to each user's preferences. For example, if a user expresses ignorance about or gives higher importance to a particular sub-topic in the pre-study survey questionnaire, higher weight can be assigned to the key terms from that sub-topic to retrieve more relevant documents. This would result in a personalized set of results for each user. However, for this user study, we decided not to change the retrieval query based on each participant's inputs. Instead, we modeled a sample dummy user who was considered to be ignorant in a few critical topics; and used the corresponding user model to assign high weights to some topics and low weights to others. This helped us control the exact set and sequence of documents that all participants would see in the user study.

Inviting participants

Volunteers were invited to participate in the study by announcing the study on mailing lists in many departments within a large, diverse, public university. Emails were sent out primarily to graduate students and staff members. Invitation emails were also sent to members of the larger community (local, but not affiliated to the university), and to participants of a nation-wide multi-center research meeting. There were no specified eligibility or language requirements for participation in the study.

The announcement invited volunteers to participate in a learning task, where they were expected to learn about a topic and answer questions. Participants were not informed of the exact nature of the study or what factors were being measured. They were, however, informed that they can participate in the learning tasks related to two topics, and that each task would take about 45 minutes to complete. They were asked to learn as much as they could about the various sub-topics within the task. They had an option to quit the study phase whenever they felt they had read enough passages. They were also informed that, as a token of appreciation for their time and participation, they would be rewarded a fixed amount for each task they successfully complete.

Volunteers who agreed to participate in the study were issued a unique identifier (pseudonym) that they would use to access the study. The pseudonyms were statically assigned to two interface variants from the ones listed in Table 1. The variants were assigned in such a way that each participant would see two fairly different interfaces for the two tasks. One of the tasks was randomly assigned to the one of the former four variants (that showed one or two documents per page), and the second task was randomly assigned to one of the latter four variants (that showed ten documents per page). Participants were allowed to select any of the two tasks first, so the researchers did not have control over which version of the interface participants saw, and for participants that took part in both topics, the order in which they saw them.

After giving explicit online consent to the study, participants would start with the pre-study survey questionnaire. All responses were recorded with the pseudonym and no identifiable information was requested or recorded during the study.

#	***Milk***	***Energy***
1.	organic milk	impact on economy/jobs
2.	raw milk	relation to climate change
3.	flavored milk in schools	increased oil drilling, coal production
4.	calcium from milk	carbon capture/clean coal technology
5.	vitamins, minerals from milk	nuclear power and safety concerns
6.	lactose intolerance	ethanol and bio-fuels
7.	early puberty in children	solar power
8.	effect on cancer, diabetes	wind power
9.	impact of dairy industry	fuel cells
10.	pus cells, added hormones	hydro power

Table 3.

Concepts covered in the survey questionnaire for each task.

Pre-study survey

Each participant was asked to choose one of the topics to begin the study. At the start, they were asked generic questions about the issue. For the *Milk* task, for example, participants were asked about their dietary preference, i.e. whether they identified themselves as vegans, lacto-vegetarians, or non-vegetarians. They were asked whether they drank milk regularly, and if so, how often, and their reasons to drink or not drink milk. They were also asked specifically if they believed milk was a healthy choice for human consumption. Similarly, participants choosing the *Energy* task were asked if they drove a gasoline-driven, electric, or hybrid car, if they biked or walked to work, or used public transportation. They were also asked specifically if they believed alternate energy sources were a viable alternative to fossil fuels.

This was followed by a series of questions on specific sub-topics relevant to the issue being studied. For the *Milk* task, these included questions to gauge the participants' knowledge and preference of conventional milk over raw milk or organic milk, flavored milk in schools, nutrients in milk such as calcium and vitamin D, lactose intolerance and milk allergies, effect of milk on early puberty and cancer, and impact of milk consumption on the economy and the environment. Similarly, for the *Energy* task, the questionnaire asked participants to think about the issue in the context of specific alternative sources of energy such as ethanol, bio-fuels, nuclear power, solar power, hydro power, wind power, and hydrogen fuel cells, and the impact of alternate energy sources on job creation and the economy. It also included questions about traditional sources of energy such as oil, coal, and natural gas, and their impact on global climate change. Table 3 lists the concepts covered in the pre-study survey questionnaire for the two issues considered in this study.

Study phase

Participants were randomly assigned to one of the interface variants described in Table 1. For each relevant passage, participants were shown the topic of the passage and whether the passage was in favor of, against, or neither in favor nor against the topic. All actions that participants took with respect to the study were logged. This included the choice and order in which passages were selected, the time taken for each passage, and the responses to questions.

A group of participants were observed while they took the study. They were asked to think aloud (verbally articulate their thoughts) while they learned about one of the controversial topics. Observational notes were recorded based on their interaction.

Once the participants completed the three online stages, they were shown a success code that denoted successful completion of the study task. They could then either participate in the second study task using the same pseudonym or quit the study.

Feedback interview

The lead researcher met with all participants after they completed the online components of the study. Typically, this was a ten minute interview, where the participants were debriefed about the factors being studied. Participants were also informed about other interfaces being studied. The participants who went through both study tasks were asked about their relative experience with the two interfaces they encountered.

Particulars	**Overall**	***Milk***	***Energy***
Number of documents read	18.6	20.1	17.1
Number of documents skipped	12.6	13.0	12.1
Time spent in study phase (in min)	26.5	26.5	26.6

Table 4.

Readership statistics of the study topics.

Analysis of Study Results

User Profile and Interaction Summary

Volunteers in the age group of 18 to 65 were invited to participate in the study. In all, 24 volunteers participated in the study, and the average age of participants was 28.6 ± 4.9 years. The group of participants consisted of eighteen males and six females. All participants had a college degree and only six participants were native speakers of English. Each participant could take part in at most two study tasks. In all, we collected information from 40 study tasks; with most participants choosing to take part in both tasks.

The profile of how participants interacted with the system was similar for both tasks. Typically, participants took 7 – 10 minutes to complete each of the pre-study and post-study questionnaires. On-an-average, they spent 26.5 minutes in the study phase. Table 4 summarizes the interaction based on number of documents read and overall time spent in the study phase. We observe that participants spent almost the same time for both tasks. On average, participants read 18.6 documents, but considered as many as 31 documents during this time frame (including documents they read and those they chose to skip). The similarity in the number of documents accessed in the two tasks shows that the topics were similar in cognitive complexity and invoked similar behavior.

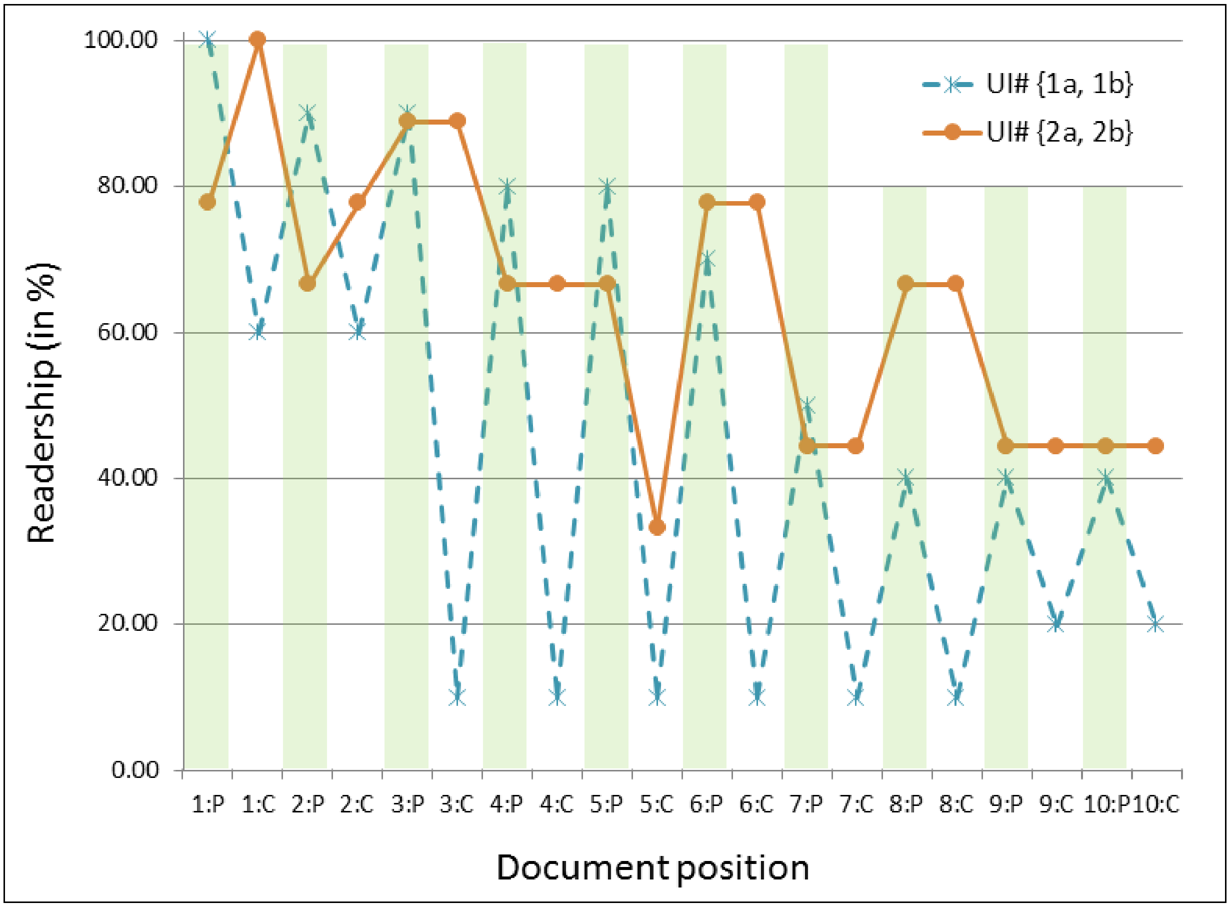


Figure 5.

Variation of number of documents read for each document position. Documents in the primary set are in odd positions (shaded green) while those in the contrast set are in even positions (shaded white).

Which documents are read and which are skipped?

Explicitly showing contrast documents helps reading both viewpoints

To understand which documents participants read and which they skip, we compared how many times participants read a document per result position. In UI variants **Single-Single-all-all** (UI# {1a, 1b}), participants are shown only one document by default. These documents belong to the primary document set. If they want, they can choose to see the corresponding contrast document next. For all other UI variants, the contrast documents are shown alongside the primary document, with documents in favor of the sub-topic on the left side of the screen, and documents against the sub-topic on the right.

Figure 5 shows how the readership changes with document position for the top 10 results. We compare two scenarios -- one in which only one primary document is shown by default (UI variants **Single-Single-all-all**, UI# {1a, 1b}) and the other where one primary and one contrast document are shown side-by-side (UI variants **Single-Contrast-all-all**, UI# {2a, 2b}). As we see from the figure, the readership for contrast documents is significantly lower when it is not shown by default, while the readership of the primary document does not change much. The readership of contrastive documents increased from 22% when only the primary document was shown by default, to 64.44% when one primary and one contrast document was shown side-by-side. In contrast, the readership for the primary document dropped slightly from 68% to 64.44%. This shows that explicit display of contrastive viewpoints leads users to read both viewpoints, when compared to presentation styles that do not present, but merely suggest existence of alternate viewpoints.

UI #	**{1a, 1b}**	**{2a, 2b}**	**{3, 4a, 4b, 5}**
UI Variant	**Single-Single-all-all**	**Single-Contrast-all-all**	**Multiple-Contrast-all-all**
Number of passages displayed	1	2	10
Number of participant study sessions	10	9	21
Number of documents read	13.5	20.3	20.2
Number of documents skipped	4.3	5.1	19.7
Time spent in study phase (in min)	26.3	29.4	25.4
Time per document read (in min)	2.38	1.77	1.43

Table 5.

Interface type influences reading pattern

Showing multiple documents per page increases readership

When multiple documents are shown per page, users tend to be more selective in what they read. Table 5 summarizes the variation in reading pattern as the number of documents is increased. We observe from the table that showing only one document per page not only significantly reduces the total number of documents read, but participants tend to also spend more time reading those documents on-an-average. By showing the contrast document alongside, participants tend to spend more time overall to read more documents. When multiple documents are shown on a page, participants are able to consider and skip many more documents in the stipulated time. When we compare single document and multi-document views, we see that although participants read the same number of documents in all, they scanned about 15 more passages in the multiple document view.

Scheme	**1 star**	**2 stars**	**3 stars**	**4 stars**	**5 stars**
Bimodal					
Single document view (UI# {1a, 1b, 2a, 2b})	76.3%		85.4%		
Multiple document view (UI# {3, 4a, 4b})	39.7%		61.3%		
Uniform					
Single document view (UI# {1a, 1b, 2a, 2b})	56.3%	77.3%	85.3%	83.2%	84.3%
Multiple document view (UI# {3, 4a, 4b})	38.2%	52.8%	60.7%	66.0%	66.9%

Table 6.

Readership increases with expertise rating.

Higher expertise rating gets higher readership

Next, we looked at the impact of trust rating on which documents are read and which are skipped. For this analysis, we distinguish the UI variants based on whether the rating scheme was bimodal or uniform-at-random. Table 6 shows the variation in documents read for these two rating schemes.

We focus on two classes of interfaces -- one in which only one document (with or without the contrast document) was shown, and the other where the five primary and the corresponding contrast documents were shown. In both cases, we find that higher rated documents are read more often than those with poor expertise ratings. This is clearly seen in the multiple-document interface, where participants have more choice in what they want to read. As we see in Table 6 (last row), only 38% of documents with a trust rating of 1 were read, compared to almost 67% of documents with 5-star rating read by the participants. Even when the documents are rated under the bimodal scheme, the higher rated documents are 22% more likely to be read than the lower rated documents. These trends follow even in the single document view, where only one document (with or without the contrasting document) is shown to the participants. It is noteworthy that in this scheme, relatively fewer documents are skipped. This is because, in the single document view, there is typically no (or just one) option – i.e. to either read or skip the shown document, without the knowledge of what the next document would be. So, a relatively larger proportion of documents is read, even if the documents have low trust ratings.

Absence of trust rating boosts readership of low-rated documents, hurts others

In our study, we had one UI variant (**Multiple-Contrast-None-Sort**, UI# 5) in which the trust ratings were not shown to the participants. We find that, under this setting, 49.8% of the shown documents were read by the participants. When compared to the figures in Table 6, we can say that unrated documents are more likely to be read than those with 1-star rating.

Which documents do participants agree with?

For each passage that participants read, they are asked if they agree with what the passage talks about. The participants were asked to judge their agreement on a four point Likert scale, ranging from (i) ‘mostly disagree’, (ii) ‘somewhat disagree’, (iii) ‘somewhat agree’, and (iv) ‘mostly agree’. Participants also had an option to reply with a ‘can't say’, if they were not sure of the response.

Scheme	**No rating**	**1 star**	**2 stars**	**3 stars**	**4 stars**	**5 stars**
Bimodal		2.91 (1.03)		2.77 (0.92)		
Uniform	2.86 (0.88)	2.46 (0.84)	2.90 (0.90)	2.83 (1.00)	2.57 (0.94)	2.92 (0.89)

Table 7.

Variation of agreement score with expertise rating. The numbers in parentheses show variance.

Expertise rating does not affect agreement

As described in the previous section, we considered the bimodal and the uniform rating schemes. Table 7 summarizes our findings on how the agreement scores vary with changes to expertise rating. We find that irrespective of the expertise rating, participants tend to agree with the documents to the similar extent (average ratings are slightly above 2.5). Although high-rated documents appear to have relatively higher agreement scores, we do not find a statistically significant correlation of the agreement scores with the variation in expertise levels. Pearson’s correlation coefficient between the agreement scores and the expertise rating under the uniform scheme was 0.448.

However, in the feedback interviews, 73% of the participants claimed that they tend to agree with highly rated documents. Further analysis of the results is required to understand this discrepancy. One possible explanation is that even low-rated documents do not give false information, and so it is hard to find instances where participants disagree with the information provided.

Another interesting deviation was that the average agreement scores for documents rated 4-star was considerably lower than when documents were rated 3-star or 5-star. This seemed counter-intuitive. When we looked at the data more deeply, we find that we had assigned one of the sources, the ‘Dairy Industry of America’, a 4-star rating in the uniform rating scheme. Participants considered this source to be a highly biased source and tend to give very low agreement scores to documents from this source. This indicates that in cases when participants clearly know the bias of the source, they tend to ignore the expertise rating; even if they otherwise trust the expertise rating. The analysis gave us additional evidence on how users interacted with the information presented to them, when it is counter-intuitive to their knowledge.

Which documents do participants find informative?

Similar to the question on agreement, for each passage read, participants are asked if they find the passage informative. The participants were asked to respond on a four point Likert scale, ranging from (i) ‘no new information at all’, (ii) ‘some new information’, (iii) ‘a lot of new information’, and (iv) ‘all new information’. Participants also had an option to reply with a ‘can't say’, if they were not sure of the response.

Scheme	**No rating**	**1 star**	**2 stars**	**3 stars**	**4 stars**	**5 stars**
Bimodal		1.87 (0.75)		2.05 (0.72)		
Uniform	1.88 (0.72)	1.40 (0.49)	1.88 (0.76)	1.84 (0.63)	1.65 (0.66)	1.97 (0.73)

Table 8.

Variation of informativeness score with expertise rating. The numbers in parentheses show variance.

Expertise rating weakly correlates to informativeness

As described in the previous section, we considered the bimodal and the uniform rating schemes. Table 8 shows a summary of how informativeness varies with expertise rating. We find that under both schemes, higher expertise rating seems to be weakly correlated to the informativeness score. Pearson's correlation coefficient between the informativeness score and the expertise rating under the uniform scheme was 0.634.

Informativeness strongly correlates to agreement

We also investigated if participants agreed or disagreed with the passages they found informative. The Pearson correlation coefficient between informativeness score and the agreement score assigned to passages was very high ($r = 0.971$, $p < 0.01$), indicating that participants strongly agreed with the passages they found informative.

Which documents do participants rate as biased?

Similar to the question on agreement, for each passage read, participants are asked if they find the passage biased, and if so, whether it was biased in favor of or against the issue at hand. However, unlike the other two measures, the participants were asked to respond to this on a five-point Likert scale, ranging from (i) 'strongly biased against', (ii) 'somewhat biased against', (iii) 'unbiased', (iv) 'somewhat biased in favor of', and (v) 'strongly biased in favor of' the issue-at-hand. As with the other questions, participants also had an option to reply with a 'can't say' if they were not sure of the response.

Scheme	No rating	1 star	2 stars	3 stars	4 stars	5 stars
Likert scale [1,5] bias averages						
Bimodal		2.62 (1.13)		2.99 (1.22)		
Uniform	3.16 (1.26)	2.93 (1.48)	3.18 (1.28)	3.36 (1.20)	3.11 (1.40)	3.30 (1.49)
Average bias strength on [1,3] scale						
Bimodal		2.08 (0.71)		2.00 (0.70)		
Uniform	2.05 (0.72)	2.27 (0.77)	2.02 (0.79)	2.02 (0.74)	2.29 (0.67)	2.35 (0.69)

Table 9.

Variation of bias score with expertise rating. The numbers in parentheses show variance.

Both highly-rated and poorly-rated documents perceived to be strongly biased

As described in the previous section, we considered the bimodal and the uniform rating schemes. Table 9 shows a summary of how the perception of bias varies with expertise rating. The top half shows the average bias score using Likert scale values as is. The bottom half of the table shows the strength of the bias on a [1, 3] scale. The strength was computed by transforming the Likert scale scores such that both strong biases for and against the topic get a score of 3, followed by both weak biases being assigned a strength score of 2, and the unbiased judgments getting a score of 1.

Using the bias strength transformation (bottom half of Table 9), we find that both very poorly rated and very highly rated documents were perceived to be strongly biased. This insight gets hidden when the Likert-scale ratings are aggregated as-is.

Extent of Learning

We next look at the learning task, and analyze what factors helped or hindered learning about the controversial topics. We were able to capture this based on participants' responses to identical questions in the pre-study and post-study questionnaires.

Participants learned about topics they did not know

In our study, participants tended to read more on topics they did not know about, rather than read about topics they already knew. In all, during the pre-study survey phase, participants indicated not knowing about a particular phenomenon or sub-topic on 86 occasions (in about 10.8% cases). Out of these 86 instances, the same participants later reported, in the post-study survey phase, to have learned something about the sub-topic in 63 instances. This constitutes a learning rate of 73.26%.

Participants changed strongly-held biases

The study also helped participants to moderate strong biases towards the issues. At the start of the *Milk* study task, participants were asked if they considered that milk was a healthy for human consumption. Most participants overwhelmingly believed it to be a very healthy choice. The average rating for milk being a healthy food choice in pre-study survey was 4.55 ± 0.59 on a five-point Likert scale. However, after being exposed to evidence about possible contamination of milk, added chemicals, and adverse impact on health for certain individuals, the average rating for milk as a healthy food choice in the post-study survey reduced to 3.91 ± 1.08. This is a statistically significant reduction in previously held belief. Many participants also noted this explicitly in the essay they were asked to write about what they learned in the study. One participant wrote: "*I did not know that milk had so many worrisome factors caused due to mass-scale production. I have to be more careful!*" On further analysis, we also found that participants who did not read many documents also did not change their biases. Only 40.5% of the participants who read less than 15 documents changed their bias. On the other hand, participants who were inquisitive and read both the documents and their contrasting viewpoints had a higher tendency to change their strongly-biased opinions by at least one point on the Likert scale. In our analysis, 64.7% of the participants who read more than 20 documents on interfaces that displayed contrastive viewpoints changed their bias.

On the second task on *Energy*, participants were asked if they believed alternate energy sources are viable alternatives to fossil fuels. On-an-average, we did not find an overwhelming bias for this issue. The average rating for this question in the pre-study survey was 2.80 ± 0.51 on a four-point Likert scale, which means most people believed that alternate energy sources can replace close to significant portion of power generated by fossil fuels. In the post-study survey, we find that the optimism increased, but only slightly. The average rating for the same question in post-study survey was 2.98 ± 0.55 on a four-point Likert scale.

Type of questions	**Measure**	**Total number of questions**	**Number of questions in which**		**Mean %age change in the measure**
			Measure increased	**Measure decreased**	
Milk					
Knowledge questions	Mean Knowledge	9	7	2	+12.3 % *
Bias questions	Spread of Neutrality	11	2	9	-31.0 % *
Energy					
Knowledge questions	Mean	13	8	5	+3.3%
Bias questions	Spread of Neutrality	7	2	5	-27.9 % *

Table 10.

Relative improvement in responses to knowledge and bias questions after the study phase. Knowledge questions are measured on the mean knowledge score, where higher values are better. Bias questions are measured on the spread of neutrality, where lower values are preferred.

Learning about sub-topics relevant to the study task

Next, we looked at individual sub-topics to see if learning improved in the sub-topics that participants read about. Participants changed their opinion about certain sub-topics in 285 instances. Out of these, in 40.8% of the cases, the participants changed the importance they gave to the sub-topics significantly (by at least one point on a four-point Likert scale). Similarly, in 24.3% cases, participants reported to have reduced the strength of bias by at least one point, moving away from extreme bias positions.

For each task, we had twenty questions that were either knowledge oriented or bias oriented about specific sub-topics. Table 3 lists the concepts covered by the survey questions; and the complete list in given in the appendix. For knowledge questions, we measured whether the participants showed evidence of increase in knowledge, and used the average knowledge rating as the aggregate measure. For bias questions, we measured whether highly biased views were moderated, and participants had a more neutral perspective about the topic after the study. We used the spread of the neutrality rating as an aggregate score, and lower values are preferred. Table 10 summarizes the relative improvement in knowledge and bias questions. In the *Milk* task, there were nine knowledge related questions, out of which seven questions got an overall higher rating and two questions got a poorer rating. The average knowledge rating increased by 12.3%. Similarly, there were eleven bias related questions in the *Milk* task. We observed that the neutrality spread rating reduced in nine of the bias questions and increased in the remaining two, with an average reduction in the measure by 31.0%. The increase in the average knowledge rating and the reduction in the bias neutrality rating are both statistically significant at p=0.05 level using Wilcoxon's signed rank test.

Similarly, for the *Energy* task, we found that the neutrality spread reduced in five out of the seven bias-related questions, with an average reduction of 27.9%, which was also a statistically significant improvement at p=0.05 level. However, although the knowledge rating increased in eight out of thirteen knowledge questions, the average increase of 3.3% was not found to be statistically significant. On further analysis of the responses, we found that many participants read only about a few sub-topics, i.e. they did not read about most alternative energy sources. So, because of limited exposure, their opinion about viability of alternative energy source replacing fossil fuels did not change significantly. However, participants with a strong bias against the issue demonstrated increased knowledge about the viability of alternate energy sources, and reduced their bias after the study phase.

Interface variant groups	**Number of times the variant groups were ranked**			**Aggregated Score**
	First	**Second**	**Third**	**(3 x 1st + 2 x 2nd + 3rd)**
Single-Single-all-all **(UI# {1a, 1b})**	2	8	8	30
Single-Contrast-all-all **(UI# {2a, 2b})**	7	6	5	38
Multiple-Contrast-all-all **(UI# {3, 4a, 4b, 5})**	9	4	5	40

Table 11.

Comparison on interface variants groups on how effectively they reduce spread of neutrality for bias questions. Scores in the last column are aggregated votes following the Borda counting strategy (Cohen, 2000), as shown.

Comparative display also helped reduce previously held biases

We looked at which interface variants helped reduce strong biases held by participants. We considered three variant groups – first, the single document view without contrastive viewpoint (UI variants **Single-Single-all-all**, UI# {1a, 1b}); second, the single document view with contrastive viewpoints (UI variants **Single-Contrast-all-all**, UI# {2a, 2b}); and third, the multiple document view with contrastive viewpoints (UI variants **Multiple-Contrast-all-all**, UI# {3, 4a, 4b, 5}). For each bias question, we compared the three interface variants groups and ranked them based on their effectiveness in reducing the spread of neutrality score.

Table 11 summarizes our findings. The last column of the table shows an aggregated score, based on weighted aggregation of votes for each variant group, following the Borda counting strategy (Cohen, 2000). We observe that the interfaces with contrastive viewpoints are ranked higher than single viewpoint variants in 16 out of 18 bias questions. Further, the interface variants with multiple document view are more effective in reducing the previously held biases than the interfaces with single view.

Participants displayed knowledge of more content words after the study

Finally, we analyzed the keywords and concepts mentioned by participants in the pre-study and post-study surveys. We find that participants mentioned more relevant keywords in the post-study survey than in the pre-study survey. Table 12 lists the content words in the top twenty most frequent words under each category. We observed that participants were more specific in identifying key terms that supported or opposed the issue being considered.

	Milk		***Energy***	
	pre-study	**post-study**	**pre-study**	**post-study**
Terms in favor	calcium vitamins source nutrition bones protein	calcium vitamins nutrients source children protein minerals healthy lactose important provides	oil pollution fossil environment limited power peak natural renewable clean	power renewable jobs nuclear alternative pollution environmental security fossil hydrogen hydro
Terms against	Cows hormones fat products lactose production pressure	pus cells cancer hormones cows lactose fat flavored unhealthy intolerance cloned sugar	cost alternative environmental expensive power economic efficient	environmental wind hydro-power efficiency production cost water

Table 12.

Content words occurring in top 20 most frequent keywords in pre- and post-study surveys.

Interface factors		**Average effectiveness rating (max 3.00)**
Displaying contrasting viewpoints side-by-side		2.58
Showing contextual information		2.21
Showing information about the viewpoint: whether the document is in favor or against the issue		2.13
Showing expertise rating	Multiple document view	2.00
	Single document view	1.55
Ordering results before display	Sorting on topic	1.71
	Sorting on relevance	1.53

Table 13.

Relative importance of interface factors to help participants in completing their task.

What interface factors helped participants?

We also asked participants to self-report which factors helped them in learning about the topic. For each UI factor, participants rated how helpful the factor was on a three-point Likert scale, corresponding to whether the factor was (i) ‘not helpful at all’, (ii) ‘somewhat helpful’, or (iii) ‘very helpful’. The findings are summarized in Table 13. The highest rating was given for the display of contrasting viewpoints side-by-side, or in the case of UI# {1a, 1b}, the ability to explore contrasting viewpoint. Overall, the participants gave it an average rating of 2.58 on the three-point Likert scale.

Participants also found context information to be very helpful. Although many participants did not know the individual sources such as authors or organizations, they could utilize the qualification of the source and the passage context (if it was from a journal, blog article, or an email) to decide if they want to read the passage. The average rating for display of contextual information was 2.21 on the three-point Likert scale. Similarly, the information about the viewpoint was also appreciated. It received an average rating of 2.13 on the same Likert scale.

Participants gave a relatively lower rating to the expertise information. Participants found the multi-document view to be somewhat helpful, with the average score of 2.00 on the three-point Likert scale. To compare, the single document view was found to be less helpful, with an average score of 1.55 on the same three-point Likert scale. Based on the feedback interviews, it appears that many participants were apprehensive about the system that generated the expertise ratings. Some participants misunderstood the rating to have come from ratings of other users and ignored them. It is interesting to observe that while participants found the expertise information to be less useful, the expertise rating did seem to have an impact on their reading pattern, as noted earlier.

Finally, participants were asked if the order of documents helped them in their task. Participants who encountered views where documents were topic sorted (UI variants **all-all-all-Sort**, UI# {4a, 4b, 5}) gave a slightly higher rating than participants who saw the documents coming in relevance sorted order. The average score for relevance-sorted interface variants was 1.53, while that for the topic-sorted variants was slightly higher at 1.71 on the three-point Likert scale.

Conclusion

Providing access to unbiased information is critical to satisfying information needs in many domains. However, for controversial claims, it is important to understand which factors affect the perception of credibility and how to overcome the human tendency to stick to one's own viewpoint. We conducted a user study called BiasTrust to understand these factors. We varied various parameters to test which factors significantly help users to learn about a controversial topic.

We find that, when compared to merely providing the option to look at documents from alternate viewpoints, showing contrasting viewpoints by default helped significantly reduce strong biases in favor of or against topics and helped participants learn about new sub-topics in an unbiased fashion. We also observe that showing expertise rating helps participants pick which documents to read and which to omit. This effect is more prominent when the sources are given very low ratings. Further, documents with expertise ratings that are very low or very high also invoke a perception of bias. So, care must be taken to justify and calibrate the ratings generated by an automated system.

Although this is an initial study on how to present controversial topics to potentially biased users, the findings are already interesting. Participants spent over an hour for each task and gave valuable feedback, explicitly during post-study evaluation and feedback sessions, and implicitly by choosing which passages to read. The insights gained by this study will help us and other researchers optimize the design of an automated claim verification system, which could not only learn which evidence documents are most relevant to show to the user, but also what additional information needs to be provided to help users assimilate the information faster.

We hoped to get a diverse set of participants for the study, but we were limited in our search. Specifically, we did not get a diverse, yet uniform representation in political views in our participant base. Of the 24 participants in our study, ten participants declared themselves as independent, and seven as politically leaning democrat. The remaining seven participants chose not to declare their political inclination. It would be interesting to also study if the interaction behavior changes with political viewpoints, especially on politically-polarized topics. The conclusions drawn based on our analysis are also limited by the scope of the user study, as it does not model the inherent variations in human personality traits, such as open-mindedness, willingness to change beliefs, potential to be swayed by persuasion, persistence of changes, and lack of topical interest in information-seeking behavior. Further studies are needed to evaluate interaction of interface design with the cognitive and psychological learning models on overcoming bias and information-seeking.

Another potential limitation of our study is that we maintained the set of documents seen by the participants as constant, in order to observe how all participants interacted with the same set (and ordering) of documents. This study, hence, analyzes a specific use case of information seeking where the assembled documents are all identified a priori as credible and relevant evidence documents. Further, the degree of controversy is also controlled and known. These design decisions introduce some selection bias in the experiments that may potentially undermine the applicability of some of the findings more generally. In practice, one would also like to study how the documents can be varied based on what the participant already knows, and measure if a more targeted set of retrieved documents would significantly improve the knowledge and bias ratings. A comprehensive evaluation of the claim verification system would address some of these aspects ignored in the current work.

This study can be further extended to understand how to effectively summarize evidence to give an overall perspective first before getting into details. A larger scale study is planned, involving much smaller tasks, to look into these aspects. Some participants also suggested simplifying the technical terms used in some of the passages to help laypersons understand the issues better. This is indeed an interesting need, but beyond the scope of the study presented in this paper.

Acknowledgments

The work presented in this paper was done when the first author was at the University of Illinois and over a summer at Palo Alto Research Center. The authors would like to thank Prof.Wai-Tat Fu and Jason Coronel for their help in the design of the user study. Some of the survey results reported here were obtained from searches of the iPOLL Databank and other resources provided by the Roper Center for Public Opinion Research, University of Connecticut. This research is partially funded by grants from CCICADA (Command, Control, and Interoperability Center for Advanced Data Analytics), a MIAS (Multimodal Information Access and Synthesis) Center of Excellence of the Department of Homeland Security (DHS) at the University of Illinois at Urbana-Champaign. The research was also partially sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on. Parts of the research and the user study were also sponsored by grants from the Information Trust Institute and the Office of Naval Research.

References

Baron, J. (2000). Thinking and deciding. Cambridge Press.

Blei, D. M., Ng, A. Y., Jordan, M. I., & Lafferty, J. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993-1022.

Croft, W. B. (2000). Combining approaches to information retrieval. In W. Bruce Croft, editor, Advances in Information Retrieval: Recent Research from the Centre for Intelligent Information Retrieval. Kluwer.

Danielson, D. R. (2005). Web credibility. In C. Ghaoui (Ed.), Encyclopedia of human-computer interaction. (pp. 713-721). Idea Group.

Fogg, B. J. (2003). Persuasive technology: Using computers to change what we think we do. Morgan Kaufmann.

Fogg, B. J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., Paul, J., Rangnekar, A., Shon, J., Swani, P., & Treinen, M. (2001). What makes Web sites credible? A report on a large quantitative study. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI) (pp. 61-68).

Fogg, B. J., & Tseng, H. (1999). The elements of computer credibility. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI) (pp. 80-87).

Gallard, A., Abiteboul, S., Marian, A., & Senellart, P. (2010). Corroborating information from disagreeing views. In Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM) (pp. 131-140). ACM.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In Proceedings of 22nd ACM International Conference on Research and development in Information Retrieval (SIGIR) (pp. 50-57). ACM.

Kaiser Family Foundation. (2004). Seniors and the Internet. Survey conducted by Kaiser Family Foundation and Princeton Survey Research Associates International, March 5-April 18, 2004. USPSRA2004-ENT031. Retrieved from the iPOLL Databank, The Roper Center for Public Opinion Research, University of Connecticut.

Kaiser Family Foundation. (2009). 2009 Survey of Americans on HIV/AIDS – Toplines. Survey conducted by Henry J. Kaiser Family Foundation, January 26-March 8, 2009. Retrieved from the iPOLL Databank, The Roper Center for Public Opinion Research, University of Connecticut.

Leiserowitz, A., Maibach, E., Roser-Renouf, C., & Hmielowski, J. D. (2011). Politics and global warming: Democrats, Republicans, Independents, and the Tea Party. Yale University and George Mason University. New Haven, CT: Yale Project on Climate Change Communication. Retrieved from http://environment.yale.edu/climate/files/PoliticsGlobalWarming2011.pdf

Lemur Project. (2002). The Lemur Toolkit. Retrieved from http://www.lemurproject.org/

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. Psychological Science for the Public Interest, 13(3), 106-131.

Metzger, M. J., Flanagin, A. J., Eyal, K., Lemus, D. R., & McCann, R. M. (2003). Credibility for the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. In P. J. Kalbfleisch (Ed.), Communications yearbook, 27, 293-335. Erlbaum.

Nickerson, R.S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2(2), 175-220.

Pariser, E. (2011). The filter bubble: What the Internet is hiding from you. Penguin Press.

Pew Research Center for the People and the Press. (2010). Americans Spending More Time Following the News – Ideological News Sources: Who Watches and Why. Pew Research Center Biennial News Consumption Survey.

Pirolli, P., Wollny, E., & Suh, B. (2009). So you know you're getting the best possible information: A tool that increases Wikipedia credibility. In Proceedings of the 27th annual SIGCHI Conference on Human factors in Computing Systems (pp. 1505-1508). ACM.

Plous, S. (1993). The psychology of judgment and decision making. McGraw-Hill.

Prior, M. (2003). Liberated viewers, polarized voters: The implications of increased media choice for democratic politics. The Good Society, 11, 10-16.

ProCon.org (2011). ProCon.org. Homepage: http://www.procon.org/

PSRA/Newsweek/Kaplan Education Center. (1998). Survey of national adult parents of children in grades K-8, conducted by Newsweek, Kaplan Educational Centers, March 5 - March 10, 1998. [USPSRNEW.031298.R16]. Retrieved from the iPOLL Databank, The Roper Center for Public Opinion Research, University of Connecticut.

PSRA/Newsweek. (2000). PSRA/Newsweek Poll # 2000-05: K-8 Parents. Survey of national adult parents of children in grades K-8, conducted by Newsweek and Princeton Survey Research Associates, February 5 - February 10, 2000 [USPSRA2000-NW05]. Retrieved from the iPOLL Databank, The Roper Center for Public Opinion Research, University of Connecticut.

Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the Web. Journal of the American Society for Information Science and Technology, 53(2), 145-161.

Rieh, S. Y., & Belkin, N. J. (1998). Understanding judgment of information quality and cognitive authority in the WWW. In Proceedings of the 61st Annual Meeting of the American Society for Information Science (ASIS) (pp. 279-289).

Rieh, S. Y., & Danielson, D. R. (2007). Credibility: A multidisciplinary framework. In B. Cronin (Ed.), Annual Review of Information Science and Technology, 41, 307-364. Information Today.

Savolainen, R. (2011). Judging the quality and credibility of information in Internet discussion forums. Journal of the American Society for Information Science and Technology, 62(7), 1243-1256.

Suh, B., Chi, E. H., Kittur, A., & Pendleton, B. A. (2008). Lifting the veil: Improving accountability and social transparency in Wikipedia with WikiDashboard. In Proceeding of the 26th annual SIGCHI Conference on Human factors in Computing Systems (pp. 1037-1040). ACM.

Taber, C. S. & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. American Journal of Political Science, 50(3), 755-769.

Taylor, H. (2011). The Growing Influence and Use of Health Care Information Obtained Online. The Harris Poll® #98, published on September 15, 2011. Harris interactive.

Ugander, J., Backstrom, L., Marlow, C., & Kleinberg, J. (2012). Structural diversity in social contagion. In Proceedings of the National Academy of Sciences of the United States of America (pp. 1-5).

Vydiswaran, V. G. V, Zhai, C. & Roth, D. (2011). Gauging the Internet doctor: Ranking medical claims based on community knowledge. In Proceedings of the KDD Workshop on Data Mining for Medicine and HealthCare (pp. 42-51).

Vydiswaran, V. G. V., Zhai, C., Roth, D., & Pirolli, P. (2012a). BiasTrust: Teaching biased users about controversial topics. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM) (pp.1205-1209). ACM.

Vydiswaran, V. G. V., Zhai, C., Roth, D., & Pirolli, P. (2012b). Unbiased learning of controversial topics. In Proceedings of the 75th Annual Meeting of the American Society for Information Science and Technology (ASIS&T) (pp.291.1-291.4).

Wathen, C. N., & Burkell, J. (2002). Believe it or not: Factors influencing credibility on the Web. Journal of the American Society for Information Science and Technology, 53(2), 134-144.

Wu, M., & Marian, A. (2007). Corroborating answers from multiple Web sources. In Proceedings of the 10th International Workshop on Web and Databases (WebDB). (pp. 1-6).

Yin, X., Han, J., & Yu, P.S. (2008). Truth discovery with multiple conflicting information providers on the Web. IEEE Transactions on Knowledge and Data Engineering, 20(6), 796-808.

Appendix

In this appendix, we list the complete list of topics included in the pre- and post-survey questionnaire, along with the classification of the question as a bias or a knowledge question.

Issue: *Milk*

1. Importance of organic milk. (knowledge)
2. Preference of organic milk over conventional milk. (bias)
3. Preference of raw milk over conventional milk. (bias)
4. Importance of flavored milk. (knowledge)
5. Opinion about flavored milk being healthy or unhealthy. (bias)
6. Importance of calcium in the diet. (knowledge)
7. Opinion on whether milk is a rich source of calcium. (bias)
8. Importance of vitamins and minerals in the diet. (knowledge)
9. Opinion about milk being a rich source for vitamin D. (bias)
10. Importance of lactose intolerance. (knowledge)
11. Importance of early onset of puberty. (knowledge)
12. Opinion about milk causing early onset of puberty? (bias)
13. Effect of milk on cancer/diabetes. (knowledge)
14. Opinion about whether consuming milk has an effect on cancer? (bias)
15. Effect of dairy industry. (knowledge)
16. Effect of advertisements. (knowledge)
17. Environmental concerns due to milk production. (bias)
18. Opinion about milk from cloned animals. (bias)
19. Opinion about artificial bovine growth hormone in milk. (bias)

20. Opinion about pus cells in milk. (bias)

Issue: *Energy*

1. Opinion about whether alternate energy sources can replace conventional energy sources. (bias)
2. Importance of alternate energy sources creating more jobs. (knowledge)
3. Opinion about government subsidies for alternate energy sources (bias)
4. Importance of increasing energy independence and security. (knowledge)
5. Importance of Global climate change due to conventional energy sources. (knowledge)
6. Importance and viability of increasing oil drilling and its impact. (knowledge)
7. Importance and viability of increased fossil fuel usage. (knowledge)
8. Importance and viability of increased natural gas usage. (knowledge)
9. Importance and viability of development of clean coal technology. (knowledge)
10. Opinion on relevance of clean coal storage technology. (bias)
11. Importance and viability of Ethanol and biofuels. (knowledge)
12. Opinion about harmful effects of Ethanol and biofuels. (bias)
13. Importance and viability of nuclear power. (knowledge)
14. Opinion about safety concerns with nuclear power. (bias)
15. Importance and viability of solar power. (knowledge)
16. Opinion about impact of solar power on the environment (bias)
17. Importance and viability of wind power viable (knowledge)
18. Opinion about impact of Wind power impact (bias)
19. Importance and viability of hydrogen fuel cells. (knowledge)
20. Importance and viability of hydro power. (knowledge)