A Crowdsourcing Methodology to Measure Algorithmic Bias in Black-box Systems: A Case Study with COVID-related Searches*

Binh Le[1][0000−0001−6429−5053], Damiano Spina[1][0000−0001−9913−433X], Falk Scholer[1][0000−0001−9094−0810], and Hui Chia[2][0000−0003−3075−3364]

[1] RMIT University, Melbourne, Australia
{binh.le, damiano.spina, falk.scholer}@rmit.edu.au
[2] The University of Melbourne, Melbourne, Australia
chia.h@unimelb.edu.au

Abstract. Commercial software systems are typically opaque with regard to their inner workings. This makes it challenging to understand the nuances of complex systems, and to study their operation, in particular in the context of fairness and bias. We explore a methodology for studying aspects of the behavior of black box systems, focusing on a commercial search engine as a case study. A crowdsourcing platform is used to collect search engine result pages for a pre-defined set of queries related to the COVID-19 pandemic, to investigate whether the returned search results vary between individuals, and whether the returned results vary for the same individual when their information need is instantiated in a positive or a negative way. We observed that crowd workers tend to obtain different search results when using positive and negative query wording of the information needs, as well as different results for the same queries depending on the country in which they reside. These results indicate that using crowdsourcing platforms to study system behavior, in a way that preserves participant privacy, is a viable approach to obtain insights into black-box systems, supporting research investigations into particular aspects of system behavior.

Keywords: Crowdsourcing · Algorithmic bias · Search engines

1 Introduction

"Should I get vaccinated for COVID-19?" is a question that many people may ask a web search engine or an intelligent assistant nowadays. Would users find the same information if they ask the same question in the negative form, e.g., "Should I avoid getting vaccinated for COVID-19?" Anyone who expects the retrieval

* This work has been partially supported by the Australian Research Council (ARC) Centre of Excellence for Automated Decision-Making and Society (ADM+S, CE200100005). Damiano Spina is the recipient of an ARC DECRA Research Fellowship (DE200100064).

system to be fair and unbiased might expect that overall, the key information that is returned would allow them to ultimately draw the same conclusion.

Beyond specific nuances of how a search query is phrased, modern web search engines combine query matching and ranking functions together with multiple signals, including a user's search history, click behavior, and location, so as to maximize the likelihood of retrieving search results or answers to satisfy that specific user's information need. As a consequence of personalizing the user experience, different individuals may get different search results for the same queries. In some scenarios, e.g., health or security-related queries, this can have undesirable implications, as different people may be exposed to information with different content, or inconsistent levels of reliability and trustworthiness. Moreover, the phenomenon of *echo chambers* – where information access systems such as search engines or recommender systems reinforce existing preferences of users in a feedback loop [1, 5, 8, 9] – is particularly problematic in this context. In this work we consider a commercial web search engine and health-related queries. This is especially relevant in the context of the COVID-19 pandemic, where health information has been highly politicized, and echo chambers of false information regarding COVID-19 have been found on social media platforms such as Twitter [7] or YouTube [2].

The aim of this study was to investigate whether individuals receive different search results for health information queries, based upon a difference in opinion as expressed in the wording of the query. For instance, if a person was more inclined to get vaccinated, they are more likely to search "Should I get vaccinated"; whereas if a person is already less inclined to get vaccinated, they may be more likely to search "Should I avoid getting vaccinated". This is within the context of online misinformation and politicization of health information surrounding the COVID-19 pandemic, where the spread of scientifically inaccurate misinformation has emerged as a risk to public health and safety [11, 13, 16].

The rationale underlying this study is that search engines have an ethical obligation to give all individuals equal access to credible health information from authoritative sources, regardless of the individual's current opinion on that health topic. The personalization of Search Engine Result Pages (SERPs) is generally acceptable for most topics, as it can improve the user experience and usually does not cause harm. However, the personalization of SERPs when it comes to critical information such as regarding health may cause direct harm to the community, if it leads to some individuals receiving less credible information. Thus, we sought to identify whether a person's current viewpoint about controversial topics related to the COVID-19 pandemic, tested by expressing the same query posed in negative or positive terms, would impact the quality of their search results from authoritative sources.

In this setting, we consider the following questions:

- Do different individuals get the same or different search results for the same queries?
- Do results vary between positive and negative query formulations for the same person?

We tested the feasibility of using a crowdsourcing methodology to quantify algorithmic bias when treating the underlying information access and retrieval system as a black-box, which is the case with commercial web search engines that keep their ranking processes as tightly controlled corporate secrets. Using Amazon Mechanical Turk, we asked 50 crowd workers to submit a set of 10 queries related to the COVID-19 pandemic – including both positive and negative forms of expressing the same information needs – to a commercial web search engine (i.e., Google) and to upload the de-identified Search Engine Results Page (SERP) that they obtained.[3]

Our results demonstrate that different individuals can indeed receive different search results for the same queries, based on factors such as the country in which the searcher is located. While this is not an unexpected result, it validates the sensitivity of the proposed method. More surprisingly, we also found that results can vary substantially between positive and negative query formulations.

2 Preliminaries

2.1 Auditing Algorithmic Bias

Friedman and Nissenbaum [4] defined a biased system as one that systematically treats specific individuals or groups differently from others, providing either unfair advantages or disadvantages. As computer applications and their development processes become more complex, the definition of bias became multifaceted. To understand a black-box system and whether its workings exhibit possible bias, a range of approaches are available. A number of studies have investigated the advantages and disadvantages of these different methods for auditing the fairness of systems [12, 14].

Five approaches for the auditing of systems, as described by Sandvig et al. [14], are:

Code Audit (Algorithm Transparency): With this method, one simply looks directly at the source code of an algorithm. However, algorithms are commonly trade secrets and highly protected by the owning company, whose competitiveness and revenue may be directly impacted by the effectiveness of their system. Furthermore, given the complexity of modern systems, it's very challenging for third parties to audit source code directly, line by line, without a tremendous amount of effort, often including needing explanations from the developer.

Noninvasive User Audit: This approach is conducted by surveying users of the platform, rather than examining the platform itself. This, however, is not easy when it comes to getting a representative sample, and the results may themselves suffer from a high degree of bias due to the limitations of human memory and emotions related to the users' experience of the platform.

[3] The data collection process for this work was reviewed and approved by RMIT University's Human Research Ethics Committee (project number 23588).

Scraping Audit: This method is more programmatic, and involves writing automated scripts that make use of a system's API services, or directly download and process system outputs (e.g. raw HTML markup from a web page). The downside of this approach is that it does not reflect the way normal users would interact with system in an everyday context. Another major challenge of this approach is legal; under the US Computer Fraud and Abuse Act (CFAA), a researcher who attempts to do this might open themselves to legal action, with penalties possibly including jail terms.

Sock Puppet Audit: Instead of involving real users, the researcher creates a software program that behaves like one. If the platform cannot distinguish between the program and a real user, this method can provide useful data.

Crowdsourced Audit / Collaborative Audit: This method differs from the previous methods in that it gets real humans to work on a task as designed by the researchers. With this method, the platform has to treat the request as if it comes from real humans, as they are real humans.

2.2 Auditing Platforms and Search Engines

In recent years, watchdog organizations such as AlgorithmWatch[4] [10, 15] or initiatives such as Ad Observer,[5] part of the NYU Cybersecurity for Democracy project, have proposed *data donation* methodologies to investigate the transparency and accountability of automated decision-making (ADM) systems deployed in online platforms such as recommender systems on Instagram or advertising engines on Facebook. These initiatives ask volunteers to donate the data they observe in a platform or a search engine by installing a plugin in their browser. The ADM+S Australian Search Experience project[6] uses a similar methodology to understand to what extent the SERPs retrieved using a common set of search queries (e.g., 'federal elections') differ across different users aged 18 or older and currently residing in Australia.

2.3 Crowdsourcing Platform: Amazon Mechanical Turk

Amazon Mechanical Turk[7] (MTurk) launched in 2005 as a crowdsourcing platform for tasks that cannot be completed by a machine and require human contribution, including identifying and characterizing objects, voices, images, etc. [12]. The system brings together *requesters* and *workers*: the former set up and publish series of work tasks (often called HITs or human intelligence tasks) to be done, and the latter browse and choose from a list of available tasks that they can complete at their convenience.

MTurk work requests can come from different countries, as can workers (i.e., participants in work tasks). However, it is known that most of the MTurk workers reside in the United States, followed by India [3, 6].

[4] https://algorithmwatch.org/en/
[5] https://adobserver.org/
[6] https://www.admscentre.org.au/searchexperience/
[7] https://www.mturk.com/

The actual demographics of particular participants in a given MTurk project may vary substantially, for a range of reasons including the nature of the work, the amount of reward (money) offered, or even the time of day at which a task is launched. Requesters can place restrictions on workers, including for example setting limits on countries, past worker performance, and so on.

2.4 Measuring Similarity among SERPs

We consider two similarity metrics to compare SERPs: Rank-Biased Overlap (RBO), which considers the order of individual search results in a SERP; and, Jaccard similarity, which compares SERPs as sets (i.e., without considering the ranking order or individual results).

Rank-Biased Overlap. Rank-biased overlap (RBO), introduced by Webber et al. [17], is a metric to quantify the similarity of two lists, with the ability to determine the relative weighting of earlier and later items in the list. RBO allows us to compare two sets of Google search results based on the websites they returned, and their order. RBO scores between two lists range from 0 to 1, with 1 indicating that the lists are identical and 0 indicating no similarity at all. The RBO function includes a parameter p that models the *persistence* of a user inspecting a SERP. In practice, the parameter p adjusts the weight given to earlier results. By default, p has a value of 1, meaning that the weights applied to the items in the list become arbitrarily flat, and the evaluation becomes arbitrarily deep, i.e., the user would inspect all the search results in the SERP. A lower value of p gives more weight to top results; when p is 0, only the highest-ranked item in the list is considered. RBO can be calibrated to an expected stopping depth $n = \frac{1}{(1-p)}$. For instance, if we want to model the scenario where the users would pay most attention to the top three search results on a SERP, we would set the stopping depth to $n = 3$, which corresponds to RBO with $p = 0.\overline{6}$.

Jaccard similarity. Jaccard similarity is a popular similarity metric to measure the overlap between two sets. Given two sets A and B, Jaccard similarity is defined as the cardinality of the intersection divided by the cardinality of the union of the two sets:

$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Both RBO and Jaccard similarity metrics range between 0 and 1, and higher scores indicate higher similarity. In our setting, a score of 0 is obtained when two SERPs have no overlap in the search results they contain, while 1 represents identical SERPs.

3 Methodology

In this section we describe the overall process used to collect data via crowdsourcing. We also detail the configuration of the crowdsourcing task, the queries included in our study, and the process to de-identify the collected data.

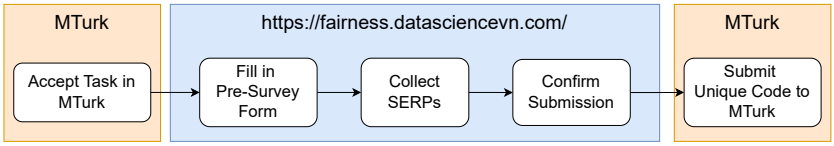

Fig. 1. Overall process used for collecting data via MTurk.

3.1 Crowdsourcing Search Engine Result Pages

The Amazon Mechanical Turk (MTurk) service was used to collect Search Engine Reult Pages (SERPs) from crowd workers. The data collection process consisted of five steps, illustrated in Figure 1.

1. **Accept the Task.** The crowdsourcing task is listed in MTurk, where workers can view a brief introduction to the task and the amount of reward obtained by completing the task. The workers then can decide if they are willing to participate. Once they start the survey, participants receive a link to our experiment.
2. **Fill in Pre-Survey Form.** Participants start by filling a pre-task questionnaire to collect demographic information such as gender, age, country of residence, and level of education. The pre-task questionnaire can be viewed at: https://fairness.datasciencevn.com/survey/start (Accessed: 21 Feb 2022) .
3. **Collect SERPs.** Participants are directed to the perform the main task of our study, requiring them to:
   (a) Manually run a provided query using Google search
   (b) Save the SERP as a HTML file, and
   (c) Upload the file into our web application.
   For data verification purposes, we configured our system to analyse the content of the HTML file and determine whether the query in the file exactly matches the query that the participant is required to upload at that stage. For example, the participant could be required to upload HTML search results for "should i not get tested for covid", however, either due to confusion or laziness, they may attempt to upload the results for a different query. In this case, the system rejects the uploaded file and asks the participant to upload again. For convenience, we created a YouTube video[8] to explain the process to participants.
   Participants repeat this process for each of the 10 search queries described below, one query at a time. We shuffle the order of queries displayed to participants, to minimize possible ordering effects.
   The HTML file submitted by the participants may contain information about the participant's Google profile, such as profile image and e-mail address, if they are signed in while submitting a query. To protect participant

[8] https://www.youtube.com/watch?v=RucW_Ok7EdQ (Accessed: 21 Feb 2022).

anonymity, as soon as each HTML file is submitted to the web application, the system automatically runs a data de-identification process: JavaScript code is run in the client side of the web application, which automatically removes all personal information (if present) and then saves the de-identified version of the HTML file. Therefore, only the de-identified version of the HTML is stored at the server-side for later analysis.

4. **Confirm Submission.** At the end of the task, the MTurk workers receive a unique code generated by our system.
5. **Submit Unique Code to MTurk.** Workers are redirected back to MTurk to submit the unique code. This code is then used to verify the submission of the worker and accept the valid tasks. Once the tasks are approved, MTurk automatically pays the workers for their completed tasks.

Crowdsourcing Setup. Participants were compensated for completing the task by a payment of US$1.20, based on the average estimated completion time of 10 minutes. In total, 50 workers were recruited to complete the task. Each worker could submit only one task. The maximum task completion time was set to 1 hour, allowing participants to take short breaks if needed, but encouraging them to focus and complete the task in a constrained block of time.

3.2 Queries

For our experiment, we devised a set of queries relating to public health considerations and beliefs around COVID-19, including testing, facemasks, and controversial treatments such as hydroxychloroquine.

Table 1 shows the queries included in our study. The queries are grouped into five semantically related pairs, one representing an information need in a *positive* form, and the other representing the information need in a *negative* form with the use of negation. We intentionally explored different ways of representing negation in queries. For Queries 2, 4, and 8 we used 'not'; for Query 6 we used 'avoid'; and for Query 10 we used the prefix 'in'.

4 Results and Discussion

4.1 Collected Data

We initially requested 50 crowdsourcing tasks to be completed. One participant submitted an incorrect validation code; therefore, only one survey needed to be republished. The average time the participants took to complete both the pre-task questionnaire and the task itself was 23 minutes 37 seconds. We launched the crowdsourcing tasks on October 5, 2020 and all the 50 valid tasks were completed by October 7, 2020.

A total of 500 SERPs (10 per participant) were obtained, resulting in a total of 4,692 items (accounting for repetition). Most of the SERPs consist of

Table 1. Queries included in our experiment, organized in pairs consisting of positive and negative expressions of an information need.

ID	Query	Pair
1	should i get tested for covid	Pair 1
2	should i not get tested for covid	
3	should i get flu shot	Pair 2
4	should i not get flu shot	
5	should i get vaccinated	Pair 3
6	should i avoid get vaccinated	
7	should i wear facemask	Pair 4
8	should i not wear facemask	
9	is hydroxychloroquine effective for covid	Pair 5
10	is hydroxychloroquine ineffective for covid	

9 or 10 items/search results. The amount of organic search results vary depending on the layout of the first page; Google search may include additional information (e.g., common questions related to COVID-19), leaving less room for organic search results. The dataset and source code used for our analysis are publicly available at https://github.com/rmit-ir/crowdsourcing-algorithmic-bias (Accessed: 21 Feb 2022).

4.2 Demographics

The crowdsourcing task was carried out by a total of 50 crowd workers residing in different countries: US (34), India (9), Brazil (5), Germany (1), and Spain (1). This is broadly in-keeping with the population of workers who use the MTurk platform [6].

Participants reported their gender as female (12), male (11), other (13), or preferred not answer this question (14). In terms of age, participants were skewed towards younger ages: 18–24 (1), 25–34 (29), 35–44 (12), 45–54 (6), 55–64 (1), and 65+ (1). For level of education, the participants reported: College degree/bachelor's degree (32); Some college (some community college, associate's degree) (7); Postgraduate or professional degree, including master's, doctorate, medical or law degree (6); Some postgraduate or professional schooling, no postgraduate degree (4); High school graduate or GED (includes technical/vocational training that does not count towards college credit) (1).

4.3 Do Different Participants Get Different Search Results for the Same Queries?

First we compared the SERPs obtained by the participants for the 10 queries included in our experiment. Given the set of participants $\mathcal{W} = \{w_1, \ldots, w_{50}\}$ and the set of queries $\mathcal{Q} = \{q_1, \ldots, q_{10}\}$, we compare the SERP for a given

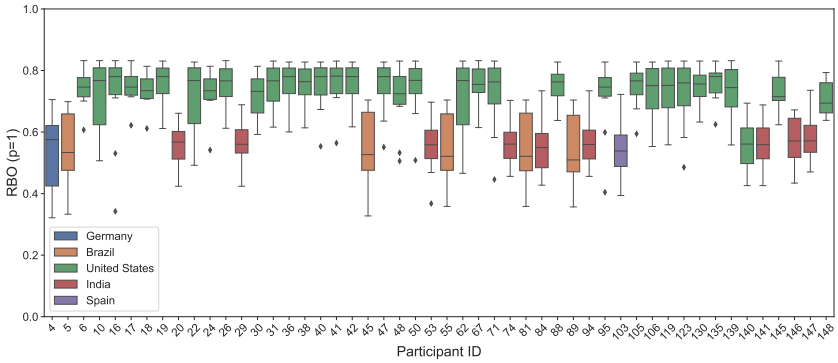

Fig. 2. Distribution of mean RBO ($p = 1$) scores, grouped by participant, when the full URL is considered to compare items in SERPs. Colors indicate the country of residence as indicated by participants.

query $q \in \mathcal{Q}$ seen by participant w_i against the SERPs seen by the rest of the participants $w_j \neq w_i \in \mathcal{W}$ for the same query q, i.e., a *between participant* analysis. We then compute the arithmetic mean of the similarity scores obtained for a given query q. This process is repeated for all queries $q \in Q$, resulting in 10 similarity scores per participant.

Individual results in SERPs can be compared at two levels: considering the full URL of the item, or considering only the domain of the item (e.g., `vic.gov.au`). Given that we have two similarity measures (RBO, and Jaccard) this results in four combinations for comparison. Figure 2 shows the results for RBO considering full URLs of items in the compared SERPs. The trends for the other three configurations where highly similar, and are not included due to space limitations.

Overall, independently of the similarity metric (RBO or Jaccard) and the granularity (full URL or domain) used, different participants may see different search results when they submit the same query to a web search engine such as Google. Even in the setting most tolerant to differences (Jaccard similarity of sets of domains and not taking rank position into account, Figure 3), we found that distribution of similarity scores cover a wide range of scores for many of the participants. Participants that indicated the country of residence as Germany, Brazil, India, or Spain, are more likely to see different SERPs that those obtained by the majority of participants residing in the US.[9] This suggests that submitting the same query may lead to different search results, depending on the location from which the query is submitted. Note that, Even though we provided a link to submit the queries to the same search engine's domain (i.e., `google.com`),

[9] The participant with ID 140 who indicated US as country of residence had substantially lower similarity scores than the other participants residing in the US. However, we do not have sufficient data to better understand the reason behind this difference, and note that the self-reported location may be inaccurate.

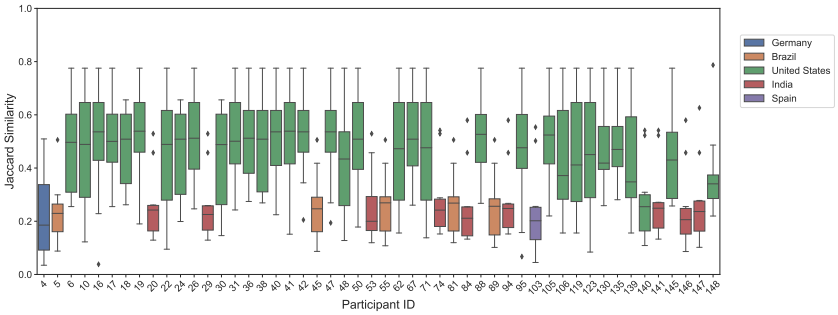

Fig. 3. Distribution of mean Jaccard similarity scores, grouped by participant, when the domain is used to compare items in SERPs. Colors indicate country of residence as indicated by participants.

the search engine automatically redirects requests to the region-specific endpoint (based on user's IP), unless a specific region is manually set in the settings.

In addition, we analyzed the results using RBO with a persistence parameter of $p = 0.67$, which corresponds to a scenario where attention is focused on the top-3 search results, and observed similar trends.

5 Do Results Vary Between Positive and Negative Query Formulations?

We performed a *within participant* analysis to investigate whether participants would see different SERPs when issuing queries in positive or negative formulations, e.g., "should i get flu shot" vs. "should i *not* get flu shot". From our data, for each query pair we have 50 similarity scores, one per participant. Figures 4 and 5 show the distribution of RBO ($p = 1$) scores when comparing the SERPs obtained by each participant for each query pairs, when full URLs and domain are considered to compare items, respectively.

When full URLs are considered in the measurements (Figure 4), RBO scores are relatively low, indicating that participants tend to obtain different answers depending on the wording used to formulate the query. At the domain level, for query pairs 1–3, the returned search results are more similar in terms of the domains they cover (Figure 5). However, this is not the case for query pairs 4 and 5, where positive and negative query formulations tend to retrieve search results from different domains. Comparing the two graphs, we can see that Query Pair 3 in particular seems to to retrieve different pages, but within common domains.

It can be seen that participants residing in the US are grouped separately compared to other countries. For query pair 2, participants residing in the US tend to get higher RBO scores compared to the rest (for both full URL and domain). However, query pairs 3 and 4 show a different trend: when full URLs are considered, participants residing in the US tend to obtain different search results with respect to the query formulation. When we look at differences at

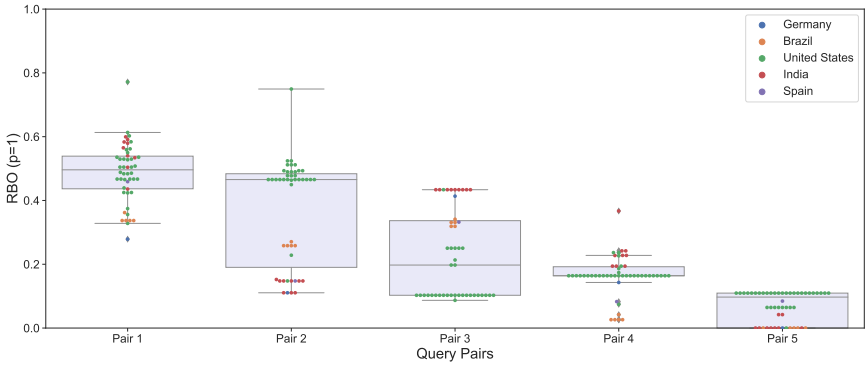

Fig. 4. RBO ($p = 1$) scores between positive and negative query formulations, grouped by pairs, when full URLs of answer items are considered. Colours represent the country of residence as indicated by participants.

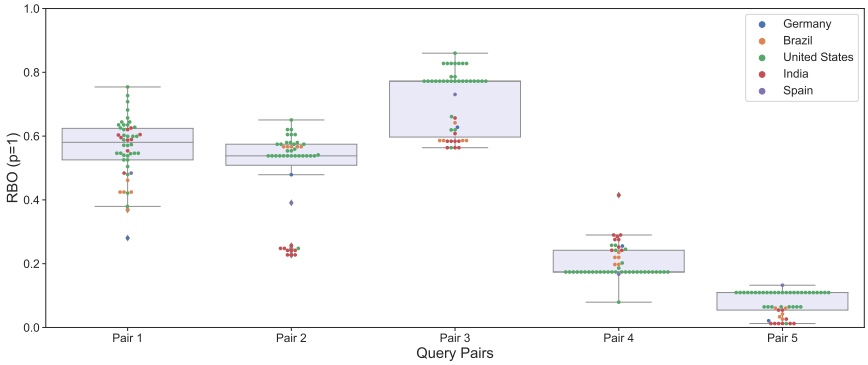

Fig. 5. RBO ($p = 1$) scores between positive and negative query formulations, grouped by pairs, when only domains of answer items are considered. Colours represent the country of residence as indicated by participants.

domain level, different query formulation leads to different search results within the same domain for Pair 3, while for Pair 4 the search results were retrieved from different domains.

We performed a similar analysis using RBO with $p = 0.67$, and observed that a larger gap appears between positive and negative query formulations. This was especially evident for Query Pair 5, where the word 'not' is not used.

Overall, our preliminary analysis indicates that, by analyzing the overlap of search results between SERPs, differences in the composition of search results are seen by people depending on whether positive or negative formulations of queries are used.

6 Conclusion

Thanks to the proliferation of the internet, information and attention are now key resources. Fair, unbiased, and ethical access to information is an underlying goal for our society. The information that we receive is largely controlled by automated technologies. However, due to the proprietary nature of commercial systems, it is difficult to measure and analyze to what extent these goals are being met.

Our pilot study validates the use of crowdsourcing platforms such as Amazon Mechanical Turk to obtain reliable data to analyze the behavior of such systems, focusing on a commercial web search engine as a case study. Our initial approach demonstrates that this can be achieved in a systematic way by developing complementary user representations: we investigated both *between user* variability, and *within user* variability, where the same information needs are instantiated using positive or negative wording. This enabled gaining insight into specific aspects of interest of black-box systems. In particular, we were able to demonstrate that the composition of search results can differ in both scenarios: different crowd workers tend to obtain different search results for the same queries, related to the country in which they are located; and, the same individuals tend to receive different search results depending on whether they use a positive or a negative query formulation.

The preliminary results with this case study validate our crowdsourcing methodology. Future work includes applying this methodology in a larger setting, including more crowd workers and more queries. We have not collected information with respect to other factors such as the browser, operating system, or device used by the crowd worker to obtain the SERP. We plan to investigate such factors in future work. We also plan to analyze the differences in SERPs in terms of information quality, and the trustworthiness of sources.

We believe that approaches such as this – in addition to other more scalable but less controlled methodologies, such as data donation initiatives – will form an essential pillar to support research practices to measure algorithmic algorithmic bias in black-box systems used to access information, such as search engines, recommender systems, and intelligent assistants.

Bibliography

[1] Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., Starnini, M.: The echo chamber effect on social media. Proceedings of the National Academy of Sciences **118**(9) (2021), https://doi.org/10.1073/pnas.2023301118

[2] Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C.M., Brugnoli, E., Schmidt, A.L., Zola, P., Zollo, F., Scala, A.: The COVID-19 social media infodemic. Scientific reports **10**(1), 1–10 (2020), https://doi.org/0.1038/s41598-020-73510-5

[3] Difallah, D., Filatova, E., Ipeirotis, P.: Demographics and dynamics of mechanical turk workers. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, p. 135–143, WSDM '18, Association for Computing Machinery, New York, NY, USA (2018), https://doi.org/10.1145/3159652.3159661

[4] Friedman, B., Nissenbaum, H.: Bias in computer systems. ACM Trans. Inf. Syst. **14**(3), 330–347 (Jul 1996), https://doi.org/10.1145/230538.230561

[5] Ge, Y., Zhao, S., Zhou, H., Pei, C., Sun, F., Ou, W., Zhang, Y.: Understanding echo chambers in e-commerce recommender systems. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, p. 2261–2270, Association for Computing Machinery, New York, NY, USA (2020), https://doi.org/10.1145/3397271.3401431

[6] Ipeirotis, P.G.: Demographics of Mechanical Turk. Tech. rep., NYU Working Paper No. CEDER-10-01 (2010), URL https://ssrn.com/abstract=1585030

[7] Jiang, J., Ren, X., Ferrara, E.: Social media polarization and echo chambers in the context of COVID-19: Case study. JMIRx Med **2**(3), e29570 (Aug 2021), ISSN 2563-6316, https://doi.org/10.2196/29570

[8] Jiang, R., Chiappa, S., Lattimore, T., György, A., Kohli, P.: Degenerate feedback loops in recommender systems. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, p. 383–390, AIES '19, Association for Computing Machinery, New York, NY, USA (2019), https://doi.org/10.1145/3306618.3314288

[9] Kitchens, B., Johnson, S.L., Gray, P.: Understanding echo chambers and filter bubbles: The impact of social media on diversification and partisan shifts in news consumption. MIS Quarterly **44**(4), 1619–1649 (2020), https://doi.org/10.25300/MISQ/2020/16371

[10] Loi, M., Spielkamp, M.: Towards accountability in the use of artificial intelligence for public administrations. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, p. 757–766, AIES '21, Association for Computing Machinery, New York, NY, USA (2021), https://doi.org/10.1145/3461702.3462631

[11] Naeem, S.B., Bhatti, R.: The COVID-19 'infodemic': A new front for information professionals. Health Information & Libraries Journal **37**(3), 233–239 (2020), https://doi.org/10.1111/hir.12311

[12] Paolacci, G., Chandler, J., Ipeirotis, P.G.: Running experiments on Amazon Mechanical Turk. Judgment and Decision Making **5**(5), 411–419 (2010), URL http://journal.sjdm.org/10/10630a/jdm10630a.pdf

[13] Saling, L.L., Mallal, D., Scholer, F., Skelton, R., Spina, D.: No one is immune to misinformation: An investigation of misinformation sharing by subscribers to a fact-checking newsletter. PLOS ONE **16**(8), 1–13 (08 2021), https://doi.org/10.1371/journal.pone.0255702

[14] Sandvig, C., Hamilton, K., Karahalios, K., Langbort, C.: Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and Discrimination: Converting Critical Concerns into Productive Inquiry **22** (2014), URL https://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sandvig.pdf

[15] Spielkamp, M.: AlgorithmWatch: What role can a watchdog organization play in ensuring algorithmic accountability? In: Transparent Data Mining for Big and Small Data, pp. 207–215, Springer (2017), https://doi.org/10.1007/978-3-319-54024-5_9

[16] Tommasel, A., Godoy, D., Zubiaga, A.: OHARS: Second workshop on Online misinformation- and Harm-Aware Recommender Systems. In: Fifteenth ACM Conference on Recommender Systems, p. 789–791, RecSys '21, Association for Computing Machinery, New York, NY, USA (2021), URL https://doi.org/10.1145/3460231.3470941

[17] Webber, W., Moffat, A., Zobel, J.: A similarity measure for indefinite rankings. ACM Trans. Inf. Syst. **28**(4) (Nov 2010), https://doi.org/10.1145/1852102.1852106