



Research Article

How alt-tech users evaluate search engines: Cause-advancing audits

Search engine audit studies—where researchers query a set of terms in one or more search engines and analyze the results—have long been instrumental in assessing the relative reliability of search engines. However, on alt-tech platforms, users often conduct a different form of search engine audit. These user audits, driven by evaluation metrics that favor search results filled with pseudoscience, White nationalism, and conspiracy theories, are what we term cause-advancing audits. We explore and characterize the evaluation strategies employed in cause-advancing audits on 4chan’s “Politically Incorrect” board. We found that Yandex, which in 2024 was purchased by a group with close ties to the Kremlin, most frequently “wins” these audits. This phenomenon represents a significant and under-researched avenue for the spread of misinformation; we call for increased attention from researchers and policymakers on the influence of smaller search engines.

Authors: Evan M. Williams (1), Kathleen M. Carley (1)

Affiliations: (1) School of Computer Science, Carnegie Mellon University, USA

How to cite: Williams, E. M., & Carley, K. M. (2025). How alt-tech users evaluate search engines: Cause-advancing audits. *Harvard Kennedy School (HKS) Misinformation Review*, 6(1).

Received: August 30th, 2024. Accepted: January 16th, 2025. Published: February 11th, 2025.

Research questions

- How do alt-tech users discuss and engage with search engines?
- What is the preferred search engine for users on 4chan’s “Politically Incorrect” /pol/ board?
- How do alt-tech users evaluate different search engines?

Essay summary

- Alt-tech users compare and evaluate search engines (both informally and via audits); users on alt-tech platforms show strong preferences for Google alternatives.
- We call these engine comparisons “cause-advancing audits” and identify four recurring strategies employed by users to evaluate search engines on 4chan’s “Politically Incorrect” (/pol/) board.
- We found that Yandex is the most frequent “winner” of these audits on 4chan’s /pol/ board.
- Cause-advancing audits constitute a clear and understudied path that allows for easy discovery of articles promoting pseudoscience, White nationalism, conspiracies, and Kremlin propaganda.

¹ A publication of the Shorenstein Center on Media, Politics and Public Policy at Harvard University, John F. Kennedy School of Government.

Implications

A user on the White nationalist forum Stormfront created a thread titled “Which search engine is most conducive to advancing our cause?”² In another thread titled “What are non-Jewish alternatives to Google?”, a commenter shared their strategy of querying “goy”—a term White supremacists often use to refer to themselves—in multiple search engines “to determine if the search engine is following Jewish propaganda.”³ The commenter then encouraged other users to perform web search using Ecosia or Yandex. A user on 4chan’s “Politically Incorrect” (/pol/) board recounted evaluating 10 different search engines on their ability to return anti-vaccine content for the query “vaccine injuries,” lamenting that Google “would just provide a bunch of fact-checker garbage results.”⁴ The user then praised Yandex and encouraged other users to search “Holocaust fake” in both Yandex and Google. We observed many users on Gab performing similar comparisons.

Each of these comments contains an informal version of a search engine audit—a type of study where researchers query a set of terms across one or more search engines and analyze the results—conducted on an alt-right platform (social networks favored by White supremacists and groups banned from traditional media). In traditional academic search engine audits, researchers often evaluate search engine result pages (SERPs) to answer questions about the accuracy, fairness, or impartiality of one or more search engines (e.g., Makhortykh et al., 2020; Urman et al., 2022).⁵ While alt-tech platform audits adopt certain elements of academic audits, like comparing search results across search engines for a set of queries, alt-tech audits tend to diverge in both goals and methods. Whereas academic audits follow systematic methodologies to understand biases in algorithms and the resulting implications for all users, alt-tech forum audits are informal, explicitly activist, and aim to identify and promote search engines that return ideologically aligned results (e.g., on 4chan’s /pol/ board, results containing racist, conspiratorial, or pseudoscientific content). We term these non-academic, often-informal evaluation and promotion strategies “cause-advancing” search engine audits.⁶ This practice mirrors the dynamics of selective trust observed in fragmented media ecosystems, where users embrace sources that reinforce beliefs and biases and reject sources that challenge them.

The fragmentation of traditional and social media ecosystems has facilitated the maintaining of large-scale *bespoke realities* where users can select or manipulate information, experiences, or beliefs to align with personal preferences or biases. This process allows users to “construct or curate a version of reality—one that can be tailored to fit whatever desires or agendas drive [them]” (DiResta, 2024, p. 41). This is an important phenomenon in explaining populist expertise, the rejection of mainstream and scientific consensus in favor of “home-grown” knowledge (Marwick & Partin, 2022). Although these dynamics have extensively been studied in the context of both social and traditional media fragmentation (Donovan et al., 2019; Figenschou & Ihlebæk, 2019), the politicization and fragmentation of information retrieval—the primary function of search engines—has received far less attention (Tripodi, 2018). Just as users can curate social media feeds and print media diets that reinforce pre-existing beliefs, they can also curate web search experiences that reinforce a given belief system. Consequently, in this study, we often observed users on /pol/ promoting search engines based on their ability to return results that promote White nationalism, pseudoscience, conspiracy theories, or medical misinformation.

² <https://www.stormfront.org/forum/t1376072/>

³ <https://www.stormfront.org/forum/t1245794/>

⁴ <https://archive.4plebs.org/pol/thread/419356626/#419361999>

⁵ The belief that systems should be guided by fairness, impartiality, and accuracy reflects a certain value system, and indeed, all algorithmic audits around social phenomenon can be viewed as activist and in furtherance of a cause (Metaxa et al., 2021).

⁶ While the audits are usually informal, some cause-advancing audits we observed approach standard academic methodologies, but with questionable query selection and evaluation metrics (e.g., Ohlan, 2024; Suede, 2022).

We found that Yandex, which was purchased in 2024 by a Russian investment fund with close ties to the Kremlin, was the most recommended search engine promoted in cause-advancing audits. Even before its purchase, recent studies have shown that Yandex has increasingly served as a tool for spreading propaganda and disinformation that aligns with Kremlin objectives (Daucé & Loveluck, 2021; Kravets et al., 2023), including overwhelmingly returning results claiming COVID-19 originated in the United States (Kravets & Toepfl, 2022). Prior research has also shown that Kremlin propaganda outlets attempt to manipulate their search rankings and that Yandex amplifies pro-Kremlin misinformation and propaganda (Kuznetsova et al., 2022; Williams & Carley, 2023). While we have only quantitatively evaluated /pol/ in this paper, we found that searching for “Yandex” on Gab, Truth Social, and even X immediately surfaces similar cause-advancing audits. On X, we observed Yandex-promoting cause-advancing audits pushed by conspiratorial accounts with hundreds of thousands of followers (e.g., The Flat Earther, 2024; The White Rabbit Podcast, 2024). Additionally, though we primarily focused on cause-advancing audits in alt-right (far right, White supremacist) communities, we note that informal audits can be conducted by users on other parts of the political spectrum. For example, we observed an account with the name “Use Yandex Search Engine for Anti-Zionist Searches” with over 85,000 followers that appropriates rhetoric from the pro-Palestinian movement to spread antisemitic conspiracies, including the promotion of The Protocols of Elders of Zion, the Great Replacement myth, and blood libel.⁷ Social media and alt-tech users that are exposed to these audits are likely at higher risk of unknowingly consuming Kremlin propaganda, which could become more extreme following Yandex’s recent change in ownership.

This finding is particularly concerning given the role of search engines in shaping behaviors and political decisions. Research has shown that search rankings can have a substantial impact on both click-through rate and political decision-making (Dean, 2022; Epstein & Robertson, 2015). Annual global surveys by Edelman since 2016 have consistently found that users trust search engines over both traditional and social media (Edelman Trust Management, 2019). Furthermore, exit-poll interviews following the 2017 Virginia gubernatorial election found that voters often used Google to research candidates, treating it as an unbiased intermediary for information, while often misinterpreting the meaning of its returned website rankings (Tripodi, 2018). Users being encouraged to seek out ideologically aligned search engines because of cause-advancing audits risks deepening political polarization, conferring undue legitimacy to misinformation, spreading conspiracies, and exposing users to state-backed propaganda.

More research and more public education on search engines are both needed. Recent academic search engine reliability audit studies have suggested that Google returns the most reliable English-language results on highly politicized topics of any search engine (Kuznetsova et al., 2022; Urman, Makhortykh, & Ulloa, 2022; Urman, Makhortykh, Ulloa, & Kulshrestha, 2022). Large-scale audits have also found little search bias in Google across political lines (Metaxa et al., 2019; Robertson et al., 2018). If Google Search were to be broken up as a consequence of new or ongoing anti-trust litigation against the company (McCabe, 2024), smaller search engines could receive more visibility. Further research is needed to understand how a decrease in Google’s market share could impact the dissemination of unreliable information to the public. As previous research has shown that search rankings can have substantial impacts on user beliefs (Epstein & Robertson, 2015; Zweig, 2017), more needs to be done to inform the public of the limitations and biases built into information retrieval systems. We call for increased attention from policymakers and academics to the accuracy, biases, and impacts of smaller search engines. Media and digital literacy education programs in as early as middle school should teach algorithmic audits to help students critically evaluate information retrieval systems alongside other sources.

⁷ <https://x.com/NicoleNonya/status/1876310361193754991>

Findings

Finding 1: Cause-advancing audits are regularly conducted on 4chan’s /pol/ board.

After pulling 250,000 mentions of 30 different search engines from 4chan’s “Politically Incorrect” /pol/ board, we identified 1,218 comments that both mentioned at least three search engines and recommended one or more search engines. Comments meeting these criteria appeared an average of 15 times per month between January 2018 and July 2024. Figure 1 displays the monthly mentions of each search engine in the 1,218 /pol/ comments in this same time period. We have noted that spikes in comparisons occurred at the onset of the COVID-19 lockdowns in 2020 and during the vaccine rollout in the spring of 2021. Much of the 2020 spike was driven by a thread titled “The Great Jewtube Exodus,” which was reposted 45 times between March 7 and March 23, 2020.⁸ In each reposting, the author encouraged others to use Yandex, Bing, Startpage, or Brave Search instead of Google. These threads repeatedly resulted in conversations about search engine quality in the comments.

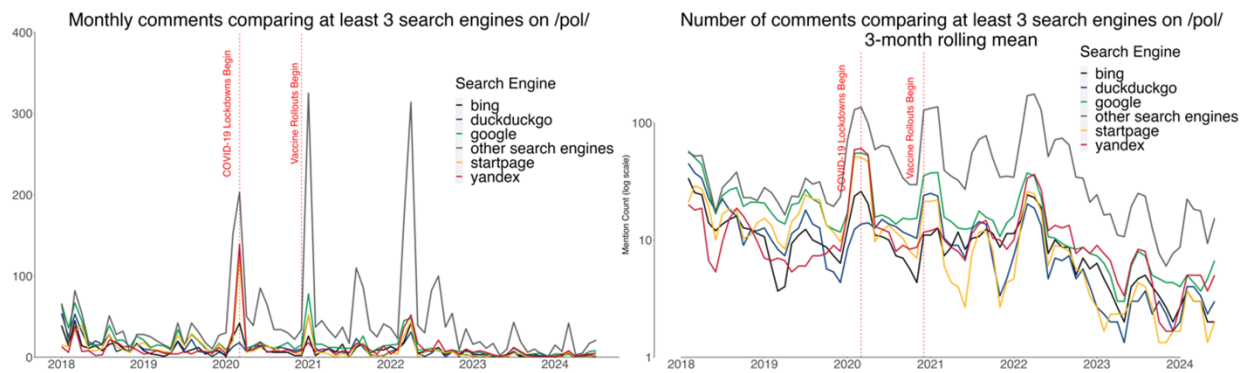


Figure 1. Search engine mention counts in 1,218 comments that mention and compare at least three search engines between Jan 2018 and July 2024 (left). The largest comparison spikes took place during the onset of COVID-19 lockdowns and the early stages of the COVID-19 vaccine rollout. For increased readability, we provided an alternative visualization (right) with a log₁₀-scale y-axis. We have applied add-one smoothing and displayed 3-month rolling means.

Finding 2: We observed that cause-advancing audits were evaluated on racial, ideological, or pseudo-scientific metrics.

Cause-advancing audits are inherently ideologically driven and evaluated. In this section, we explore several approaches that we qualitatively observed users employ when evaluating cause-advancing audits, along with specific keywords used. Many comments included no reasoning or rationale and simply promoted one or more search engines, while others offered reasoning. We identified four often-overlapping evaluation strategies: *racial counting*, *specific URL retrieval*, *engine maligning*, and *Search Engine Result Page (SERP) ideological agreement*. Quantifying and further codifying the usage of these strategies could be a fruitful path for future research.

Racial counting is an image-based evaluation strategy. In this type of evaluation, users search a term—for example, “American inventors,”—in multiple search engines and count the number of non-White individuals in the returned images. The engine that returns the fewest non-White results is then promoted by the commenter. We observed users conducting audit evaluations with this strategy for the queries

⁸ Each thread contains a different thread ID.

“White American family,” “European art,” “European people art,” “White woman,” “White couple,” “blond women porn,” and others. The inverse scenario was also used; we observed several instances of users querying racial slurs and favoring the search engines that returned the most offensive racial caricatures in image results.

Specific URL retrieval involves a user searching a specific website, article, or video (sometimes pornographic) and then discussing the experience of attempting to find the URL in multiple different search engines. The evaluation is largely binary—either the URL is found or it isn’t, but how highly it ranks is often also considered. We observed comments using this strategy for queries like “Byrne vs Clinton Foundation RICO court,” “Christchurch manifesto,” “Dontell Jackson we thought they were White,” “site:bitchute.com Hitler,” and “The original great replacement manuscript.” Some users conducting searches in this category offered more general search tips, such as using “site:” or “AROUND” operators.

Engine maligning involves disparaging a search engine—most often without clear evidence or reasoning—and using this as justification to switch to alternative engines. These are often antisemitic in nature, and users on /pol/ frequently argue that Google and DuckDuckGo are owned by Jewish interests or people. The phrase “Jewgle” appeared in 42 unique comments. Comments also veered into homophobia, with one stating that Google is “the gay nightclub of the internet[;] the only things you find there are gay and shitty.” While some comments touched on privacy concerns, these were also often steeped in antisemitism: one comment recommending the privacy-centered search engines Searx and Qwant reminded users that there is “no excuse to be feeding a Jew your search history.” Most comments that we read did not outline specific events or issues justifying ire towards the maligned search engines.

SERP ideological agreement involves quantifying how many results align with the user’s ideology. Some alt-tech media users construct formal evaluation metrics and queries. There were several comments on /pol/ that explicitly used COVID-related terms to conduct these ideological agreement audits with queries like “Ivermectin sellers,” “MRNA gene therapy,” “Pfizer vaccine,” and “Covid vax side effects.” A Gab user performed a thorough search engine comparison on Substack to find the “least censored” search engine, evaluating search engines on their ability to return proof of election fraud in the United States, Russian state media sources, anti-vaccine content, among other topics; the author found Yandex was the “winner” when evaluating on these metrics (Suede, 2022). Some users on /pol/ also approach this from the opposite direction—using the presence of Anti-Defamation League or fact-checker URLs in top results as an indicator of poor search engine quality. One commenter angrily complained that top results for “pizzagate” are all “f*ing artificial f*ing ‘debunking’ bs” and recommended that others “use Bing or DuckDuckGo for real results.”

Finding 3: Yandex “wins” the most cause-advancing audits.

We used GPT-4o-mini to extract the relative rankings of search engines in each of the 1,218 comments that compared three or more search engines. These were explicit endorsements or criticisms contained in a single comment. We found that Yandex was recommended as the top choice more than any other search engine (Table 1). DuckDuckGo and Google received the most criticism, much of which stemmed from antisemitic complaints surrounding their ownership. We included a third column in Table 1 containing the difference between the number of times a search engine was top-ranked and the number of times it was bottom-ranked. Yandex won on this metric as well.

Table 1. The number of comments that say a search engine is ranked as best, ranked as worst, and the difference between the two (best-worst difference).

Search Engine	#Comments Ranked as Best	#Comments Ranked as Worst	Best-Worst Difference
Yandex	322	213	109
DuckDuckGo	259	324	-65
Searx	226	181	45
Startpage	223	223	0
Bing	171	235	-64
Qwant	159	149	10
Google	139	298	-159

Note. We have only included statistics for the seven top-ranking search engines; the full table can be found in the released data and code.

Methods

We iteratively extracted all mentions of 30 different search engines⁹ using the `fouRplebsAPI` (Buehling, 2022). Before using the API, we slightly modified several of the search engine names in the source Wikipedia list, for example, “Yahoo! Search” to “Yahoo.” All files and operations discussed in this section can be found with the released code and data. This collection process resulted in 250,000 comments or posts mentioning at least one search engine. The data contained substantial noise, so following minor cleaning operations, we elected to only explore the set of comments mentioning at least three search engines, as these very often contained content relevant to our research questions. This resulted in a set of 1,244 unique comments between January 2018 and July 2024.

Through OpenAI’s API we provided a prompt containing instructions as well as three comments and their associated annotations (i.e., few-shot learning) to GPT-4o-mini (GPT-4) and asked GPT-4 to annotate each additional comment. GPT-4 was instructed to extract 1) whether a user was recommending one or more search engines, 2) what relative ranking (within the comment) the user assigned to each search engine—where ties should receive identical ranks, and 3) any queries that the user claims to have searched in one or more search engine. Three example comment and annotation pairs were provided with the prompt, which can all be found in the accompanying code.¹⁰ As we had qualitatively observed beforehand, GPT-4 classified most of these comments (1,177) as engine comparisons. As some of these comments were re-posted on multiple threads, our final dataset contained 1,218 comments. All results presented in this paper were conducted on that set of 1,218 comments. User evaluation approaches were determined qualitatively after reading over 400 randomly sampled comments.

To validate GPT-4’s annotations, two human annotators independently annotated a random sample of 51 of the 1,244 considered unique comments. On task 1, identifying whether a comment was an engine comparison, all annotators’ labels aligned on 42 of the 51 comments (Fleiss’ Kappa = 0.40), with 40 of the comments being agreed-upon engine comparisons. However, much of the disagreement came from human annotators; there were only two cases where the human annotators agreed and GPT4 disagreed (Fleiss’ Kappa = 0.70). We discuss prompt creation and annotation challenges in more depth in the Appendix.

⁹ https://web.archive.org/web/20240822142927/https://en.wikipedia.org/wiki/List_of_search_engines

¹⁰ <https://doi.org/10.7910/DVN/8RH7D2>

Task 2 was the most challenging of the three tasks, as it involved extracting the mentioned search engines and expressing their relative rankings. Ties were extremely common, so to measure inter-annotator agreement for the entire task, we would need to measure the similarity of three indeterminate ranked lists of sets (see Appendix for discussion and metrics definitions). However, as we were most interested in predictions at the lowest (i.e. most recommended) rank, we simply evaluated inter-annotator set similarity at rank 1 for each annotator. We find that the average Jaccard similarity of the set of rank 1 result(s) in each annotated engine-comparing comment was 0.86 for the two human annotators. Annotator 1's mean Jaccard similarity with GPT-4 at rank 1 was 0.65, and Annotator 2's mean Jaccard similarity with GPT-4 at rank 1 was 0.71. On Task 3, GPT-4 was less accurate, treating Annotator 1 as ground truth. GPT-4 correctly identified three of the five queries but also returned three false negatives with a precision of 0.5 and recall of 0.6 (see Appendix for definitions). Task 3 annotations were only to help qualitatively identify evaluation strategies, and no quantitative metrics were derived or reported from task 3.

Bibliography

- Buehling, K. (2022). *FouRplebsAPI: R package for accessing 4chan posts via the 4plebs.org API*. Zenodo. <https://doi.org/10.5281/zenodo.6637440>
- DiResta, R. (2024). *Invisible rulers: The people who turn lies into reality*. PublicAffairs.
- Donovan, J., Lewis, B., & Friedberg, B. (2019). Parallel ports: Sociotechnical change from the alt-right to alt-tech. In M. Fielitz & N. Thurston (Eds.), *Post-digital cultures of the far right* (pp. 49–65). transcript Verlag. <https://doi.org/10.1515/9783839446706-004>
- Daucé, F., & Loveluck, B. (2021). Codes of conduct for algorithmic news recommendation: The Yandex.News controversy in Russia. *First Monday*, 26(5). <https://doi.org/10.5210/fm.v26i5.11708>
- Dean, B. (2022). *We analyzed 4 million Google search results: Here's what we learned about organic click through rate*. Backlinko. <https://backlinko.com/google-ctr-stats>
- Edelman Trust Management. (2019). *Edelman trust barometer global report*. <https://www.edelman.com/trust/2019-trust-barometer>
- Epstein, R., & Robertson, R. E. (2015). The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33), E4512–E4521. <https://doi.org/10.1073/pnas.1419828112>
- Figenschou, T. U., & Ihlebæk, K. A. (2019). Challenging journalistic authority: Media criticism in far-right alternative media. *Journalism Studies*, 20(9), 1221–1237. <https://doi.org/10.1080/1461670x.2018.1500868>
- The Flat Earther [@TheFlatEartherr]. (2024, June 16). *I was using DuckDuckGo to find some information but got nowhere. Then I switched to the Russian search engine Yandex...*[Post]. X. <https://x.com/TheFlatEartherr/status/1802492973642342516>
- Kuznetsova, E., Makhortykh, M., Sydorova, M., Urman, A., Vitulano, I., & Stolze, M. (2024). *Algorithmically curated lies: How search engines handle misinformation about US biolabs in Ukraine*. arXiv. <https://arxiv.org/abs/2401.13832>
- Kravets, D., & Toepfl, F. (2022). Gauging reference and source bias over time: How Russia's partially state-controlled search engine Yandex mediated an anti-regime protest event. *Information, Communication & Society*, 25(15), 2207–2223. <https://doi.org/10.1080/1369118x.2021.1933563>

- Kravets, D., Ryzhova, A., Toepfl, F., & Beseler, A. (2023). Different platforms, different plots? The Kremlin-controlled search engine Yandex as a resource for Russia's informational influence in Belarus during the COVID-19 pandemic. *Journalism*, 24(12), 2762–2780. <https://doi.org/10.1177/14648849231157845>
- Makhortykh, M., Urman, A., & Ulloa, R. (2020). How search engines disseminate information about COVID-19 and why they should do better. *Harvard Kennedy School (HKS) Misinformation Review*, 1(3). <https://doi.org/10.37016/mr-2020-017>
- Marwick, A. E., & Partin, W. C. (2022). Constructing alternative facts: Populist expertise and the QAnon conspiracy. *New Media & Society*, 26(5), 2535–2555. <https://doi.org/10.1177/14614448221090201>
- McCabe, D. (2024, August 5). 'Google is a monopolist,' judge rules in landmark antitrust case. *The New York Times*. <https://www.nytimes.com/2024/08/05/technology/google-antitrust-ruling.html>
- Metaxa, D., Park, J. S., Landay, J. A., & Hancock, J. (2019). Search media and elections: A longitudinal investigation of political search results. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1–17. <https://doi.org/10.1145/3359231>
- Metaxa, D., Park, J. S., Robertson, R. E., Karahalios, K., Wilson, C., Hancock, J., & Sandvig, C. (2021). Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human-Computer Interaction*, 14(4), 272–344. <https://doi.org/10.1561/1100000083>
- Ohlan, T. (2024, October 9). *Google does it again, buries US-based right-leaning news on election 23-pages deep*. MRC Free Speech America. <https://mrcfreespeechamerica.org/blogs/free-speech/tom-olahan/2024/10/09/google-does-it-again-buries-us-based-right-leaning-news>
- Robertson, R. E., Jiang, S., Joseph, K., Friedland, L., Lazer, D., & Wilson, C. (2018). Auditing partisan audience bias within google search. *Proceedings of the ACM on human-computer interaction*, 2(CSCW), 1–22. <https://doi.org/10.1145/3274417>
- Suede, M. (2022). Search engine censorship test results: Find out which search engine is the least censored. *Substack*. <https://web.archive.org/web/20240224115718/https://michaelsuede.substack.com/p/search-engine-censorship-test-results?s=r>
- Tripodi, F. (2018). *Searching for alternative facts*. Data & Society. <https://datasociety.net/library/searching-for-alternative-facts/>
- Urman, A., Makhortykh, M., & Ulloa, R. (2022). The matter of chance: Auditing web search results related to the 2020 U.S. presidential primary elections across six search engines. *Social Science Computer Review*, 40(5), 1323–1339. <https://doi.org/10.1177/08944393211006863>
- Urman, A., Makhortykh, M., Ulloa, R., & Kulshrestha, J. (2022). Where the earth is flat and 9/11 is an inside job: A comparative algorithm audit of conspiratorial information in web search results. *Telematics and Informatics*, 72, 101860. <https://doi.org/10.1016/j.tele.2022.101860>
- The White Rabbit Podcast [@AllBiteNoBark88]. (2024, April 25). *Ladies & Gentlemen... A REVELATION!! have just installed Russia's Yandex App (Russian Google) See below a search on CHEMTRAILS...* [Post]. X. <https://x.com/AllBiteNoBark88/status/1783646944201736255>
- Zweig, K. A. (2017). *Watching the watchers: Epstein and Robertson's "Search Engine Manipulation Effect."* Algorithm Watch. <https://algorithmwatch.org/en/watching-the-watchers-epstein-and-robertsons-search-engine-manipulation-effect/>

Funding

The research for this paper was supported in part by the Office of Naval Research (ONR), MURI: Persuasion, Identity, & Morality in Social-Cyber Environments under grant N000142112749, ARMY Scalable Technologies for Social Cybersecurity under grant W911NF20D0002, and by the Knight Foundation. It was also supported by the Center for Informed Democracy and Social-cybersecurity (IDeaS) and the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. The views and conclusions are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ONR or the U.S. Government.

Competing interests

The authors declare no competing interests.

Ethics

All analyses were made with publicly available data.

Copyright

This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are properly credited.

Data availability

All materials needed to replicate this study are available via the Harvard Dataverse:

<https://doi.org/10.7910/DVN/8RH7D2>

Appendix: Notes on annotation

Prompt creation was an iterative process, where we iteratively added example comments and procedures to address GPT-4o-mini (GPT-4) errors on a sample of unseen comments. We first included several offensive comments containing hate-speech, to ensure that GPT4 would be able to correctly annotate them. To ensure that our few-shot examples were not biasing results, we tried annotating all data with two alternative prompts that had some examples swapped out for different ones. In the second prompt, we added two additional annotated comments that contain edge cases where we saw the model struggle—ties and responding to “green text.”

In annotations 1 and 2, we noticed that GPT4 often failed to map “Jewgle” to “Google,” so we created a third prompt where we both swapped out several of the original annotated comments and standardized engine names. In all three prompts, Yandex was by far the top-ranked search engine (min = 321, max = 331). The other top seven rankings were largely the same, but in the second prompt, Searx and Startpage had swapped rankings. All prompts and annotations are released in supplementary materials. As both annotations for prompts 2 and 3 had lower Jaccard similarity inter-annotator agreement for first-place rankings, we elected to base findings on the annotations for our original prompt. Much of the disagreement can be traced back to inherent ambiguity in the nature of the task and comments.

Extracting relative preference ranks was a challenging task both because comments could contain substantial ambiguity in relative rankings and because we were discretizing a continuous space. On the former point, for example, some comments recommend using different search engines for different functions or topics (e.g., using Bing for image search but Searx for web queries). Additionally, as we were extracting relative ranks within a comment, what “worst” means could be ambiguous. For example, if a user were to recommend Yandex over Bing, whether that user says Bing is “okay” or “the worst,” Bing still receives a relative rank of 2 under our annotation schema.

This annotation approach also presented a challenge in terms of calculating inter-annotator agreement, as a normal engine ranking for a single comment could look something like: {“yandex”:1, “searx”:1, “bing”:2, “startpage”:2, “google”:3}. This prohibits the use of standard inter-annotator agreement metrics, and even indeterminate ranked list comparison metrics like Rank-Biased Overlap fail as ranks are shared by elements. Our interest is in relative rankings, and there can be disagreement about both rank and whether an engine is being compared. As our primary research question concerns which engine is most promoted, we elected to evaluate inter-annotator agreement using the Jaccard similarity of engines each annotator assigned a rank of 1, i.e., the set of engines users most strongly recommended in each comment. The Jaccard similarity between two sets is simply the cardinality of the intersection divided by the cardinality of the union. We additionally report the precision and recall of GPT-4 on query identification (Task 3) while treating annotator 1 as ground truth. Precision and recall are defined as:

$$\textit{Precision} = \frac{\# \textit{True Positives}}{\# \textit{True Positives} + \# \textit{True Negatives}}$$

$$\textit{Recall} = \frac{\# \textit{True Positives}}{\# \textit{True Positives} + \# \textit{False Negatives}}$$