Investigating the impact of personalization on misinformation present in algorithmically curated content in YouTube

Prerna Juneja and Tanushree Mitra
University of Washington
Seattle, USA
{prerna79,tmitra}@uw.edu

ABSTRACT

Search engines are the primary gateways of information. Yet, they do not take into account the credibility of search results. There is a growing concern that YouTube has been promoting and recommending misinformative content for certain search topics. In this study, we audit YouTube to verify those claims. Our audit experiments investigate whether personalization (based on age, gender, geolocation, or watch history) contributes to amplifying misinformation. After shortlisting five popular topics known to contain misinformative content and compiling associated search queries, we conduct two sets of audits—*Search*- and *Watch-misinformative audits.* We find that demographics, such as, gender, age, and geolocation do not have a significant effect on amplifying misinformation in returned search results for users with brand new accounts. On the other hand, once a user develops a watch history, these attributes do affect the extent of misinformation recommended to them. Further analyses reveal a filter bubble effect in recommendations for all topics, except *vaccine controversies*; for these topics, watching videos that promote misinformation leads to more misinformative video recommendations. In conclusion, YouTube still has a long way to go to mitigate misinformation on its platform.

ACM Reference Format:
Prerna Juneja and Tanushree Mitra. 2020. Investigating the impact of personalization on misinformation present in algorithmically curated content in YouTube. In *Proceedings of FAccTRec Workshop on Responsible Recommendation at Recys (FAccTRec'20)* , 6 pages.

1 INTRODUCTION

Search engines are an indispensable part of our lives. Despite their importance in selecting, ranking, and recommending what information is considered most relevant for us, there is no guarantee that the information is credible. Numerous scholars have emphasized the need for systematic statistical investigations, or audits of search systems so as to uncover societally problematic behavior [13]. For example, multiple studies have audited search engines for the presence of partisan bias [9, 12] and gender bias [3, 6]. Yet, none have empirically audited them for misinformation. Moreover, investigation of video search engines, like YouTube is rare (work by Jiang et al. is one exception [10]), despite popular prediction that by 2022, 82% of internet traffic will come from videos [4]. YouTube has also faced years of criticism for surfacing misinformative content [2, 7, 15]. Critics have gone as far as calling YouTube a *conspiracy ecosystem* [1]. Despite such criticisms, there has been little effort towards quantifying the extent of misinformation in video search platforms, or investigating user attributes that might have an effect.

Our work is guided by the following main research question: What is the effect of personalization (based on age, gender, geolocation, or watch history) on the amount of misinformation presented to users on YouTube in its three major components: *search results*, *Up-Next*, and *Top 5* video recommendations? We study the conspiracy facet of misinformation and perform our audits on trending misinformative topics that are widely known to be false. In particular, we examine five misinformative topics namely, *9/11 conspiracy theories*, *chemtrail conspiracy theory*, *flat earth*, *moon landing conspiracy theories* and *vaccine controversies*. We conduct two sets of audit experiments—*Search* and *Watch* audits to examine YouTube's search and recommendation algorithms, respectively. While *Search* audits are conducted using brand new user accounts, *Watch* audits examine user accounts that have built watch history by systematically watching either all promoting, neutral, or debunking videos of potentially misinformative topics. We create over 150 Google accounts to audit YouTube. Our experiments collect 56,475 YouTube videos, spread across five misinformative topics and three YouTube components.

Guided by our main research question, we formulate following sub-questions to investigate the effects of each of the personalization attributes and in the process shed light on the phenomenon of algorithmically surfaced misinformation on YouTube.

RQ1 [*Search & Watch Experiments*]: What is the effect of demographics (age, gender) and geolocation on the amount of misinformation returned in various YouTube components?

> RQ1a [*Search Experiments*]: How are *search results* affected for brand new accounts?
> RQ1b [*Watch Experiments*]: How are *search results*, *Up-Next*, and *Top 5* recommendations affected, given accounts have a watch history?

RQ2 [*Watch Experiments*]: What is the effect of watch history on the stance of misinformative content returned in various YouTube components?

RQ3 [*Search & Watch Experiments*]: How does the amount of misinformative content differ across misinformative topics?

> RQ3a [*Search Experiments*]: How does misinformative content present in *search results* of brand new accounts differ across topics?
> RQ3b [*Watch Experiments*]: How does misinformative content present in *search results*, *Up-Next*, and *Top 5* recommendations of accounts having a watch history differ across topics?

We find little evidence to support that users' age, gender and geolocation play any significant role in amplifying misinformation in search results or recommended videos for brand new accounts.

FAccTRec'20, September, 2020
2020. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00

On the other hand, watch history exerts a significant effect on the amount of misinformation present in the *search results* corresponding to the *vaccine controversy* topic. Watch history also significantly affects the extent of misinformation in recommended videos (both *Up-Next* and *Top 5*) for all five misinformative topics. Interestingly, we observe a filter bubble effect in recommendations, where watching promoting misinformative videos lead to more promoting videos in the *Up-Next* and *Top 5* video recommendations. This filter bubble effect for recommended content is observed for all topics, except *vaccines controversies*. For the vaccine topic, while filter bubble is not observed for the *recommended* videos, it exists for the *search results*. Specifically, people who watch anti-vaccination videos are presented with less misinformation in their recommendations but more misinformation in their search results, compared to those who watch neutral or debunking vaccine videos.

2 METHODOLOGY

We present methodology for compiling high impact misinformative queries, design and implementation of our audit experiments and qualitative coding scheme for determining stance of the returned videos.

2.1 Compiling High Impact Topics and Queries

We curate a list of relevant misinformative topics by referring to Wikipedia pages on conspiracy theories [16, 17] (e.g., 9/11, chemtrail, pizzagate conspiracy, etc.). From this list, we exclude topics whose "truth" value is uncertain, that is, topics for which we were either unable to determine the mainstream perspective or the mainstream perspective is not backed by authoritative voice or scientific research. Next, we leverage Google Trends to identify the most popular topics (continuously trending, high interest topics) that are searched on YouTube by a large number of people. We discarded topics for which no trends data was returned. Then, we compare the *interest over time* plots for all remaining search topics from January 1, 2016 to December 31, 2018 and select the top 5 topics which represent the most searched topics, resulting in our list of highly impactful misinformative topics.

Next step is to generate search queries for all search topics which we use in our subsequent audit experiments and data collection. We feed seed queries representing each search topic in both YouTube and Trends and collect top 10 suggested search queries and autocomplete suggestions. Next, we manually removed duplicates and replaced semantically similar queries with a single relevant query. In total, we had 49 queries. Table 1 presents selected misinformative search topics and a few sample search queries.

2.2 Overview of Audit Experiments

YouTube utilizes age, gender, geolocation, and watch history as features in its recommendation system [5]. To determine if these features amplify the amount of conspiratorial content returned to users, we conduct a series of four audit experiments. Our audits collect three primary YouTube components namely, *search results*, *Up-Next* video and *Top 5* recommended videos on the right of the video page. We annotate the collected videos with stance values: promoting, debunking, or neutral stance towards the topic. Finally, we conduct statistical comparison tests on the annotated data.

Search Topic	Seed Query	Hot	Cold	Sample Search Query
9/11 conspiracy theories	9/11 and 9/11 conspiracy	Maryland	Ohio	9/11 inside job 9/11 tribute 9/11 conspiracy
Chemtrail conspiracy theory	chemtrail	Montana	New Jersey	chemtrail chemtrail flu chemtrail pilot
Flat Earth	flat earth	Montana	New Jersey	flat earth proof is the earth flat
Moon landing conspiracy theories	moon landing	Ohio	Georgia	moon moon hoax moon landing china
Vaccine controversies	vaccines	Montana	South Carolina	anti vaccine vaccines vaccines revealed

Table 1: Seed query, hot & cold regions, and sample search queries for the five misinformation search topics.

Our audit experiments control for multiple sources of noise. Following prior search engine audit work [8], we control for browser noise by selecting one single version of Firefox browser for all experiments. All interactions with YouTube happened in incognito mode to remove any noise resulting from tracked cookies or browsing history. We also control for temporal effects by performing simultaneous searches. Additionally, all machines used in our experiments had same architecture and version of operating system.

2.2.1 Search Experiments: Auditing with brand new accounts. For our *Search* experiments, we conduct two experiments to test whether demographics (age and gender) and geolocation for a new user (with no prior history) have significant effect on proportion of misinformative content returned by the platform.

Experiment 1: Search & Demographics (age and gender).

We consider four age groups (less than 18 years old, 18 − 34, 35-50, and greater than 50) and two gender values (male and female) (see Table 2). We create eight different Google accounts—2 (gender values) X 4 (age group values)—each having a unique combination of gender and age. We manually crafted these accounts and added appropriate profile details (age and gender).

Implementation: Each account is managed by selenium bot. The bot runs on a virtual machine created on Google Cloud Platform (GCP). When testing for demographics, searches across all accounts are performed from the same location (Mountain View, California) to control for effect of geolocation. Each bot opens Firefox browser in incognito mode and logs in to YouTube. Then it conducts searches by drawing queries from the query sets of all misinformative topics. The searches are done in sequence similar to Vincent et al's approach in [14]. The bot sleeps for 20 minutes after every search to neutralize the carry-over effect—noise introduced in search results from dependency present in consecutive searches. We collect Search Engine

Experiment #	Category	Feature	Tested Values
Search (Exp 1)	Demographics	Age	<18, 18-34, 35-50, >50
		Gender	Male, Female
Search (Exp 2)	Geolocation	IP Address	GA,MT,NJ,OH,SC
Watch (Exp 3)	Demographics	Age	<18, 18-34, 35-50, >50
		Gender	Male, Female
	Watch history	Watch history	Promoting, Neutral, Debunking
Watch (Exp 4)	Geolocation	IP Address	GA,MT,NJ,OH,SC
	Watch history	Watch history	Promoting, Neutral, Debunking

Table 2: List of user features for our audit experiments.

Annotation Value	Stance Description	Annotation Heuristics	No.of videos	Normalized Score	Sample Videos
					Video Title (Video URL, youtu.be/)
-1	debunking, mocking, disproving related misinformation	narrative of video disputes, mocks or provides authoritative evidence against conspiracy theories related to the topic under audit	430	-1 (D)	Bill Maher Throws Out 9/11 Conspiracy Theorists On Live TV (p80hXaM4QgU)
0	neutral & related to misinformation	narrative of the video does not take any stance on conspiracy theories related to the topic under audit	238	0 (N)	The Howard Stern Show and WCBS-2 On Sept. 11 (O3LT6FMF2f8)
1	promoting, supporting, justifying, explaining related misinformation	narrative of video promotes, supports or substantiates any conspiratorial views related to the topic under audit	374	1 (P)	9/11 truthers attend Treason in America (2-7GCs-2NUg)
2	debunking, mocking, disproving unrelated misinformation	narrative of video debunks, mocks or provides evidence against a conspiratorial view related to a topic different than the one under audit	64	-1 (D)	Did the Titanic Really Sink? The Olympic Switch Theory Debunked (_mpLRCqQ620)
3	neutral & related to another misinformation	narrative of the video does not take any stance on conspiracy theories unrelated to the topic under audit	25	0 (N)	JFK coverage 12:30pm-1:40pm 11/22/63 (pDOojsg62O0)
4	promoting, supporting, justifying, explaining unrelated misinformation	narrative of the video promotes, supports, justifies or explains any conspiratorial view unrelated to the topic under audit	66	-1 (P)	Mafia Boss Tells All - Jimmy Hoffa, JFK Assassination and Much More (__LxwaAEaL8)
5	not about misinformation	video content does not contain any conspiratorial views	1667	0 (N)	Former Abortionist Dr. Levatino At Virginia Tech (dIRcw45n9RU)
6	foreign language	video content in non-English language	35	translated & re-annotated	Las voces del 11S, documental en Español del Canal National Geographic (7rMQu2B_3vU)
7	undefined/unknown	annotators were unable to assign any of the above annotation values to the video	9	ignored	Ahmed Mohamed's Dad Pushes 9/11 Conspiracy Theories Online (CTkE0Etkszc)
8	removed	video removed from the platform at the time of annotation	35	ignored	n/a (tpSO7i70LHw)

Table 3: Description of the annotation scale and heuristics along with sample YouTube videos corresponding to each annotation value. We map our 9-point annotation scale to 3-point normalized scores with values -1 (Promoting, (P)) , 0 (Neutral, (N)) and 1 (Debunking, (D)). We have shared the list of 2,943 unique videos along with their annotation values in our online dataset.[1]

Results Page (SERP) for each of the 49 search queries and extract URLs of the top 20 videos.

Experiment 2: Search & Geolocation

To study the effect of geolocation, we need to identify physical locations corresponding to each search topic from where automated YouTube searches will be performed. We make use of Google Trend's *interest by sub-region* feature to shortlist locations that have the highest (hot region) or lowest (cold region) interest corresponding to each topic under audit investigation. We select one hot and one cold sub-region for each search topic based on its availability on the list of active working nodes in geographically dispersed machines, called Planet-Lab [11]. Table 1 shows the selected hot and cold sub-regions across all topics.

Implementation: For each search topic, we run two selenium bots, each corresponding to a hot or cold geolocation. These bots connect to Planet-Lab machines deployed in the hot and cold regions for that misinformative topic through ssh tunneling. After searching every query, bot saves the SERP. Later, we scrape all SERPs and extract top 20 video URLs. After completion of both search experiments (demographics and geolocation), we collected 848 unique videos.

2.2.2 Watch Experiments. The goal of our *Watch* experiments is to examine the effect that user's watch history exerts on the amount of misinformation presented to the user in both YouTube's search and video pages. The experimental setup comprises of two phases, 1) *watch* and 2) *search*. The watch phase builds the watch history of every Google account followed by the search phase that conducts searches on YouTube.

Experiment 3: Watch & Demographics. The aim of this experiment is to test the effects in the presence of user's watch history. We build history of new user accounts by automatically making them watch videos that are either all debunking, neutral or promoting the particular misinformative topic under audit investigation. We create three sets of 2 (gender values) X 4 (age group values) Google accounts to audit each misinformative topic where each set watches 20 videos from each of the three stances. We select 20 most popular videos for each of the misinformative topics. Popularity is calculated as the engagement accumulated by the video at the time of our experimental runs. It is the sum of view, like, dislike, favourite and comment count received by the video.

Implementation: Our *Watch* experiment for studying the effects of demographics is similar to our *Search* experiment runs. The only difference being that accounts build their watch history by watching, in its entirety, 20 popular videos from a particular stance set (all having the same stance in a set) before conducting any search operation on YouTube.

Experiment 4: Watch & Geolocation. The aim of this experiment is to test the effect of hot and cold geolocations on the amount of misinformation presented to the users in YouTube, given that each user has a watch history. Similar to the previous *Watch* experiment, the history is created by making each account watch YouTube videos of a particular stance. We create three sets of two Google accounts, each corresponding to a hot or cold region. The three sets build their watch histories following the same steps as in experiment 3.

Implementation: For each search topic, we run six selenium bots, three for hot and three for cold geolocations. After building their watch histories, the bot runs in a similar fashion as experiment 2—*Search & Geolocation.*

[1] https://social-comp.github.io/YouTubeAudit-data/

Feature	Topic	Stance	Comp.	Test	Mean Diff
Age	Flat Earth	N	Top5	KW H(3, 800)=18.28, p=0.0004	50 or older <all other age groups (post-hoc)
	Vaccines controv.	N	Top5	KW H(3,799)=24.65, p=1.8e-05	age 18-34 <all other age groups (post-hoc)
Gender	Flat Earth	N	Top5	MW U=74659, p=0.004	M >F
				MW U=3612, p=6.6e-07	M (50 or older) > F (50 or older)
	Moon Landing	N	Up-Next	MW U=2720, p=0.03	F >M
	Vaccines controv.	N	Top5	MW U=4068, p=0.002	M (age 35-50) > F (age 35-50)
				MW U=76206.5, p=0.02	M >F
		P	Top5	MW U=4443, p=0.01	M (age 18-34) > F (age 18-34)
			Up-Next	MW U=2880, p=0.04	M >F
				MW U=120, p=0.002	M (age 18-34) > F (age 18-34)
Geo-location	Moon Landing	P	Top5	MW U=4137.5, p=0.02	Hot >Cold

Table 4: RQ1b:*Watch* experiment results for demographics and geolocations, given accounts have built watch history after watching promoting (P), neutral (N) or debunking (D) videos. Mean corresponds to normalized scores for the annotated videos. Higher values indicate that accounts receive more promoting videos. For example, M (50 or older) >F (50 or older) indicates that males who are 50 or older and who watch neutral *flat earth* videos receive more promoting videos in their *Top 5* than females of the same age group.

2.3 Annotating our Data Collection

Through our audit experiments, we collected a total of 56,475 videos with 2,943 unique videos. We used an iteratively developed qualitative coding scheme to label our video collection. The process resulted in a scale comprising 9 different annotation values: −1 to 7. This 9-point scale gives a microscopic view of the kinds of videos a user is exposed to when she searches for a misinformative topic. For example, the videos could either promote, discuss or debunk the misinformative topic being searched, or it could discuss a different misinformative topic—a topic that the user never searched for.

Table 3 enlists annotation values with description and examples. For downstream analysis, we map our 9-point granular scale to a 3-point normalized score with values of −1, 0, and 1. The normalization process puts videos that contain any type of misinformation, whether related or unrelated to the searched topic, under the same bucket. Annotation values of 2, 3, and 4 are mapped to -1, 0, and 1, respectively, while 5 and 6 are treated as neutral (see Table 3). We discard videos coded as 7 and 8, since annotators were either unable to identify their stance (value: 7) or the video was removed from the platform (value: 8). In total, we annotated 2,943 unique videos with 501, 1980, and 462 videos marked as -1, 0, and 1.

3 RESULTS

In this section, we analyze our collected and annotated audit data to investigate our research questions. Recall that, among the three YouTube components (*search results*, *Up-Next*, and *Top 5* recommendations), we can only collect *search results* for *Search* experiments. On the other hand, we collect all three components for *Watch* experiments. A test of normality reveals that our data is not normally distributed and our samples have unequal sizes. Hence, we opt for non-parametric tests. For all pairwise comparisons, we use Mann-Whitney U test. To perform multiple comparisons, we use Kruskal Wallis ANOVA followed by post-hoc Tukey HSD[2].

3.1 RQ1: Effect of demographics & geolocation

In the first research question, we investigate the effect of demographics and geolocation on the amount of misinformation returned in various YouTube components for both brand new accounts and accounts that have build their watch history.

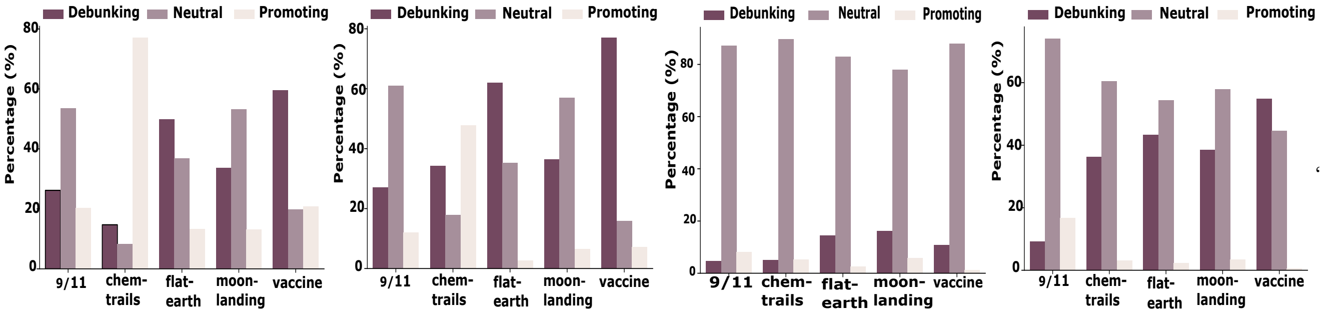
RQ1a [*Search* experiments]: How are *search results* affected for brand new accounts? We find no significant effect for gender (Mann-Whitney U = 7247667.0, p>0.48), age (Kruskal Wallis H(3,7616) = 0.00888, p>0.99), and geolocation (Mann-Whitney U=471803.0, p>0.496) demonstrating that demographics and geolocation do not have an impact on the amount of misinformation returned in search results for new users.

RQ1b [*Watch* experiments]: How are *search results*, *Up-Next*, and *Top 5* recommendations affected, given accounts have a watch history? We find that age has a significant effect for only two comparisons (refer Table 4), whereas gender has a significant effect for five comparisons involving certain combinations of search topics, watch stance, and YouTube components. In all but one significant comparisons, men receive more misinformation than females. For example, male accounts who watch neutral vaccination videos receive more promoting videos in their *Top 5* recommendations than female accounts that watch the same videos. Geo-location has a significant effect only for the *Top 5* recommendations of *moon landing* topic. Table 4 presents all the significant results.

3.2 RQ2: Effect of watch history

Watch history has a significant effect on the amount of misinformation present in *search results* of only *vaccine controversies* topic (Kruskal Wallis H(2,6517)=6.2953, p=0.0429). Post-hoc tests reveal that accounts that watch promoting anti-vaccination videos receive more promoting videos in their search results compared to those who watch neutral or debunking vaccination videos. Watch history also has significant effects on stance of misinformative videos presented in *Top 5* (Kruskal Wallis H(2,14740)=9.4235, p=0.0089) and *Up-Next* video recommendations (Kruskal Wallis H(2,2963)=10.2932, p=0.00581) when all topics are considered together. Post-hoc tests show that accounts that watch promoting videos receive more promoting results in both *Up-Next* and *Top 5* compared to those who watch either neutral or debunking videos. The effect of watch history for both these components is significant for all topics individually too. We discuss the post-hoc test results for *vaccine controversies*. Accounts that watch promoting anti-vaccination videos receive more debunking videos in their *Top 5* (Kruskal Wallis H(2,2999)=48.54, p=2.9e-11) and *Up-Next* (Kruskal Wallis H(2,600)=66.86, p=3.0e-15) components. This finding can be attributed to YouTube's initiative to reduce the recommendations of anti-vaccination videos. It is important to note that while recommendations of such videos have decreased, a filter bubble still exists with respect to the *search results*—people who watch promoting anti-vaccination videos were

[2] Tukey HSD adjusts p-values automatically, thus controlling family-wise error rate for multiple comparisons.



(a) Search Results (Search)

(b) Search Results (Watch)

(c) Top 5 (Watch)

(d) Up-Next (Watch)

Figure 1: RQ3: Percentages of video stances for each topic.

Component	Topic	Test	Mean Diff (post-hoc)
Search Results	Vaccines controv.	KW H(2,6517)=6.2953, p=0.04	P >N & P >D
Top5	All	KW H(2,14740)=9.42, p=0.009	P >N & P >D
	9/11 consp.	KW H(2,2911)=186.68, p=2.9e-41	P >N & P >D
	Chemtrails consp.	KW H(2,2845)=73.20, p=1.31e-16	P >N & N >D
	Flat Earth	KW H(2,2980)=49.18, p=2.18e-11	N >P & D >P
	Moon Landing consp.	KW H(2,3005)=17.18, p=0.0002	P >N & D >N
	Vaccines controv.	KW H(2,2999)=48.54, p=2.9e-11	N >P & D >P
Up-Next	All	KW H(2,2963)=10.29, p=0.006	P >N
	9/11 consp. theories	KW H(2,487)=60.12, p=8.8e-14	P >N & P >D
	Chemtrails consp.	KW H(2,570)=16.12, p=0.0003	P >D
	Flat Earth	KW H(2,600)=26.29, p=1.96e-06	P >D & D >N
	Moon Landing consp.	KW (2,606)=5.99, p=0.049	D >N
	Vaccines controv.	KW H(2,600)=66.86, p=3.0e-15	D >N >P

Table 5: RQ2: Analyzing watch history effects on the three YouTube components. P, N, and D are means of the normalized scores of videos presented (via the YouTube components) to accounts that have built their watch histories by viewing promoting (P), neutral (N), and debunking (D) videos, respectively. For example, P > N indicates that accounts that watched promoting videos received more misinformation (or more promoting videos) compared to accounts that watched neutral videos.

presented with more promoting content. Table 5 lists the results for the remaining topic comparisons.

3.3 RQ3: Across topic differences

In RQ3 we investigate whether misinformative content presented to users differ across misinformative topics.

RQ3a [*Search* experiments]: How does misinformative content present in *search results* of brand new accounts differ across topics? Figure 1a shows proportion of promoting, neutral, and debunking videos across all topics in *Search* experiments. We find that misinformation significantly differs among topics (Kruskal Wallis H(4,1943)=467.29, p < 7.9e-100). Post-hoc comparisons reveal that *chemtrail conspiracy theory* topic harbors significantly more misinformative *search results* compared to all other topics. Figure 1a also demonstrates the largest amount of promoting videos in the *chemtrails* topic.

RQ3b [*Watch* experiments]: How does misinformative content present in *search results*, *Up-Next*, and *Top 5* recommendations of accounts having a watch history differ across topics? Figure 1b, 1d and 1c show the proportion of promoting, neutral, and debunking videos across all topics collected from *search results*, *Up-Next* and *Top 5* recommendations respectively in *Watch* experiments. Statistical test shows that topics have a significant effect on the amount of misinformation present in *search results*, *Up-Next* (Kruskal Wallis H(4,2963)=375, p < 6.7e-80), and *Top 5* recommended videos (Kruskal Wallis H(4,14740)=390.6, p < 2.9e-83). Post-hoc comparisons using Tukey HSD reveal that *chemtrail conspiracy theories* has significantly more misinformation in its *search results* compared to all other topics (also observable from Figure 1b). On the other hand, the amount of misinformation present in *Up-Next* and *Top 5* recommendations for *9/11 conspiracy theory* topic is significantly more than other topics. This is also evident from Figures 1c and 1d.

4 DISCUSSION AND CONCLUSION

In this study, we conducted two sets of audit experiments on YouTube platform to empirically determine the effect of personalization attributes on the amount of misinformation prevalent in YouTube searches and recommendations. We found that personalization affects the amount of misinformation in recommendations once the user develops a watch history indicating a misinformation bias in the recommendations. Complete eradication of misinformation bias from YouTube recommendations requires time and significant resources. In the interim, YouTube can take several steps to tackle the problem of misinformation on its platform. It can begin by giving priority to monitoring certain misinformative topics that have a wider negative impact on society. Our work itself suggests a technique to curate such misinformative topics that are perennial, popular, and searched by a large number of people. Misinformative content belonging to the selected impactful topics can be filtered, fact-checked, and accordingly censored from the platform.

Our audits also suggest variability in YouTube's behavior towards certain misinformative topics—an indication of a reactive strategy of dealing with misinformation. We recommend the platform to also proactively reveal the workings of its algorithm. For example, users can be told "you are recommended video A because you viewed videos C and D".

Overall, our audit methodologies can be used for investigating other search engines for misinformative search results and recommendations. We believe such audit studies will inform the need for building search engines that retrieve and present results ranked according to both relevance and credibility.

REFERENCES

[1] Jonathan Albright. 2018. *UnTrue Tube – YouTube's Conspiracy Ecosystem*. https://datajournalismawards.org/projects/untrue-tube-youtubes-conspiracy-ecosystem/

[2] Nick Carne. 2019. 'Conspiracies' dominate YouTube climate modification videos. (2019). https://cosmosmagazine.com/social-sciences/conspiracies-dominate-youtube-climate-modification-videos

[3] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the Impact of Gender on Rank in Resume Search Engines. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, Article 651, 651:1–651:14 pages. https://doi.org/10.1145/3173574.3174225

[4] Cisco. 2019. Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper. (2019). https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html

[5] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*. https://doi.org/10.1145/2959100.2959190

[6] Nicholas Diakopoulos, Daniel Trielli, Jennifer Stark, and Sean Mussenden. 2018. I Vote For– How Search Informs Our Choice of Candidate. *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple, M. Moore and D. Tambini (Eds.)* 22 (2018). https://www.academia.edu/37432634/I_Vote_For_How_Search_Informs_Our_Choice_of_Candidate

[7] Renee Diresta. 2018. *The Complexity of Simply Searching for Medical Advice*. https://www.wired.com/story/the-complexity-of-simply-searching-for-medical-advice/

[8] Aniko Hannak, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. 2013. Measuring Personalization of Web Search. In *Proceedings of the 22Nd International Conference on World Wide Web (WWW '13)*. ACM, 527–538. https://doi.org/10.1145/2488388.2488435

[9] Desheng Hu, Shan Jiang, Ronald E. Robertson, and Christo Wilson. 2019. Auditing the Partisanship of Google Search Snippets. In *The World Wide Web Conference (WWW '19)*. ACM, 693–704. https://doi.org/10.1145/3308558.3313654

[10] Shan Jiang, Ronald E Robertson, and Christo Wilson. 2019. Bias Misperceived: The Role of Partisanship and Misinformation in YouTube Comment Moderation. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 278–289.

[11] Larry Peterson, Tom Anderson, David Culler, and Timothy Roscoe. 2003. A Blueprint for Introducing Disruptive Technology into the Internet. *SIGCOMM Computer Communication Review* 33, 1 (2003), 59–64. https://doi.org/10.1145/774763.774772

[12] Ronald E. Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing Partisan Audience Bias Within Google Search. *Proceedings of ACM on Human Computer Interaction* 2, CSCW (2018), 148:1–148:22. https://doi.org/10.1145/3274417

[13] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 22 (2014). https://pdfs.semanticscholar.org/b722/7cbd34766655dea10d0437ab10df3a127396.pdf

[14] Nicholas Vincent, Isaac Johnson, Patrick Sheehan, and Brent Hecht. 2019. Measuring the Importance of User-Generated Content to Search Engines. *Proceedings of the International AAAI Conference on Web and Social Media* 13, 01 (2019), 505–516. https://www.aaai.org/ojs/index.php/ICWSM/article/download/3248/3116/

[15] Cale Guthrie Weissman. 2019. Despite recent crackdown, YouTube still promotes plenty of conspiracies. (2019). https://www.fastcompany.com/90307451/despite-recent-crackdown-youtube-still-promotes-plenty-of-conspiracies

[16] Wikipedia. 2002. Conspiracy theory. (2002). https://en.wikipedia.org/wiki/Conspiracy_theory

[17] Wikipedia. 2003. List of conspiracy theories. (2003). https://en.wikipedia.org/wiki/List_of_conspiracy_theories