YouTube and Conspiracy Theories: A Longitudinal Audit of Information Panels

Lillie Godinez
Wellesley College
Wellesley, MA, USA
lg105@wellesley.edu

Eni Mustafaraj
Wellesley College
Wellesley, MA, USA
emustafaraj@wellesley.edu

ABSTRACT

One way that YouTube aims to support its users to make sense of problematic content they find on its platform is through the inclusion of information panels (IPs) in its search results or video watch pages. Such panels, when present, offer topical context from third-party sources such as Encyclopedia Britannica or Wikipedia, but also fact-checks or breaking news. What search queries trigger these IPs? Do they appear consistently over time and space? Using 620 query phrases related to conspiracy theories, crowd-sourced from 309 individuals in the United States, we conducted a longitudinal audit of YouTube search pages, simultaneously searching for the query phrases, using computers scattered in 14 locations during four months. We find that only 16.63% of the search pages contained an IP, with 114 unique query phrases triggering a total of 49 unique panels of three different types (context, fact-checks, and news). Qualitative analysis of queries and IPs reveals that keywords in query phrases match with keywords in content labels. Context IPs cover multiple queries related to the same conspiracy theory, while fact-check and news IPs are usually targeted to one particular query. However, many topical queries do not trigger an IP, indicating inconsistent coverage. Since some IPs seem to be sensitive to real-world events, more research into methods for "news-cycle-aware auditing" might be necessary.

KEYWORDS

YouTube, conspiracy theories, longitudinal audit, content labeling

ACM Reference Format:
Lillie Godinez and Eni Mustafaraj. 2024. YouTube and Conspiracy Theories: A Longitudinal Audit of Information Panels. In *35th ACM Conference on Hypertext and Social Media (HT '24), September 10–13, 2024, Poznan, Poland.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3648188.3675128

1 INTRODUCTION

Long before the social-media-fueled spreading of COVID-19 misinformation and conspiracy theories became an "urgent threat" to public health [7] and disinformation regarding the 2020 US Presidential election results led to a riot in Washington, D.C. on January 6th, 2021 [10], YouTube was home to old conspiracy theories from "the earth is flat" [27] to the "fake moon landing" [9], and served as an amplifier for new types of conspiracies that deny the truth of post-2000 events such 9/11 [31] or the Sandy Hook shooting [29]. When this tendency to provide conspiratorial explanations to certain kinds of events (like mass shootings) was repeated in the context of the fatal Marjorie Stoneman Douglas high school shooting,[1] with YouTube videos claiming that the shooting survivors were "crisis actors", YouTube came under public pressure to deal with its promotion of conspiracy theories [8].

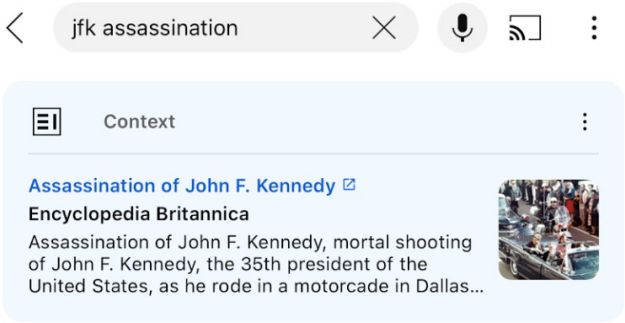

Figure 1: The screenshot of a context information panel that shows below the search query box on the YouTube search results page, for the query phrase *jfk assassination*.

One month after this episode, YouTube announced that one way it has chosen to deal with the issues of misinformation and conspiracy theories spreading on its platform was to introduce contextual information panels (IP), an instance of content labeling [23], which are implemented as small boxes of information that direct users to third-party information providers, as the example in Figure 1 shows.

In the years ever since, YouTube has continued to expand the types of IPs shown to users. Currently, when a user searches YouTube, they may see IPs that show contextual information on a topic from sources like Wikipedia, Encyclopedia Britannica, etc. (context IP); check the truth value of a particular claim (fact-check IP, refer to Figure 2); display media sources reporting on breaking news (news IP); or show information on election integrity, candidates, and voting (election IP). Meanwhile, such IPs and others also show under individual videos on their watch pages.

This continuous expansion of IPs serves as an indicator that YouTube may regard them as an important part of its strategy to deal with problematic content. Given YouTube's significant impact

HT '24, September 10–13, 2024, Poznan, Poland

ACM ISBN 979-8-4007-0595-3/24/09
https://doi.org/10.1145/3648188.3675128

[1] Also known as the Parkland high school shooting, https://en.wikipedia.org/wiki/Parkland_high_school_shooting

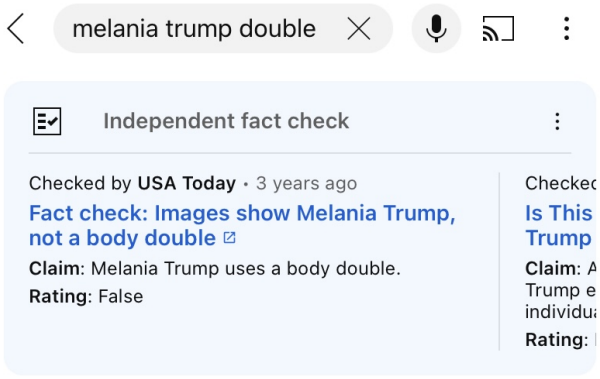

Figure 2: The screenshot of a fact-check IP triggered by the query *melania trump double*. These IPs may contain multiple fact-checks from different organizations (partially shown).

on the information environment,[2] we consider it important to study the extent, consistency, and coverage of such an intervention.

Like other social media platforms, YouTube is often emphasizing its efforts for dealing with problematic content and IPs are an appealing alternative to content moderation, avoiding outright removal of such content. However, YouTube is vague about the consistent application of IPs. Its official documentation[3] notes that:

> When you search or watch videos related to topics prone to misinformation, such as the moon landing, you **may** see an IP at the top of your search results or under a video you're watching.

(Note: The verb "may" was boldfaced by the authors to highlight YouTube's uncertainty about the presence of the panels.)

Although researchers have often audited YouTube's algorithms for its tendency to promote misinformation [13, 35], audits that focus on information panels are very rare and not comprehensive. Therefore, our study aims to examine the extent, consistency, and coverage of IPs by conducting a longitudinal audit of YouTube search pages over a four month period, in 14 locations, using crowdsourced search queries. Our investigation focuses on answering these three specific research questions, with respect to the implementation of the IP feature on YouTube:

- **RQ1**: How prevalent are IPs on YouTube search pages for query phrases about conspiratorial topics?
- **RQ2**: Does the presence of information panels (IPs) change across time and location?
- **RQ3**: What is the relationship between IP types and search query phrases?

However, since YouTube is a "living system" that is continuously changing in response to the changing nature of its content and the behavior of other actors, one snapshot of auditing (even a longitudinal one, like ours) is not going to be sufficient to answer our research questions once and for all. Thus, in addition to understanding YouTube's behavior with respect to the IPs, we are also interested in perfecting the auditing instrument by experimenting with the variables of the auditing process (queries, time, location), to understand how each of them affects the results of the audit.

Main findings and contributions: We gathered a dataset of 138,880 search engine results pages (SERPs) covering 620 queries (about conspiracy theories) over 14 locations and four months. Our analysis of these SERPs indicates that IPs are present in 16.63% of total SERP instances. Context IPs are the most common type, covering around 15% of all SERPs in the dataset, with the other two types showing significantly less often. Qualitative analysis of the queries that invoke each type of IP indicates that all IPs are triggered by keywords in the query, and a context IP tends to cover multiple queries on the same conspiracy theory. However, the presence of the keywords doesn't guarantee that a context IP is triggered, a puzzling result that needs to be further researched. Meanwhile, query phrases that invoke fact-check and news IPs are more tailored, with multiple keywords in the query phrase matching keywords in the IP, and in-line with YouTube's description of how these IPs work. Finally, we find that the IPs are more consistent across location (geographic locations) than time. Our findings have implications for how future audits need to be designed and for informing discussions about content labeling of problematic information.

2 RELATED WORK

2.1 Problematic information on YouTube

With their widespread popularity, social media platforms are particularly vulnerable to the spread of disinformation, misinformation, and conspiracy theories, collectively referred as *problematic information* [14]. Several studies have named YouTube as one of the major platforms prone to problematic information. For example, [26] describes YouTube as a favorite source for prominent online spreaders of disinformation on Twitter. Research suggests that individuals who use YouTube for news are more likely to believe conspiracy theories [36] and less likely to be fully vaccinated against COVID-19 [5]. [3] found that YouTube users who believe in conspiracy theories related to the alleged 2020 US presidential election fraud had videos endorsing such claims in their recommended videos.

Past research has tracked the prevalence of conspiratorial videos after a major YouTube policy was implemented to remove content from the platform [6], understanding disinformation campaigns on YouTube [12], and examining how user activity differs for conspiratorial content [25]. Furthermore, researchers have focused on misinformation present for well-known topics like climate change [1], COVID-19 [5, 22, 26], and the 2020 US presidential election [3].

The listed examples indicate that research on problematic information on YouTube has focused on a few popular topics that capture everyone's attention. In this study, our goal is to go beyond popular topics, taking a human-centered approach, and looking at all conspiracy theory topics people show an interest in, both popular and obscure, to determine the extent to which YouTube has implemented IPs to address them. While we agree that some conspiracy theories are more harmful than others, when taken all together, they contribute in polluting the online information environment.

[2]YouTube ranks 2nd, in terms of worldwide popularity, with 2.5 billion monthly visitors: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.
[3]https://support.google.com/youtube/answer/9004474?hl=en&ref_topic=9257092&sjid=3128292846590594282-NA, Accessed 07/29/24.

2.2 Auditing algorithmic platforms

Auditing algorithms has many uses [2], but one of the ways auditing has been used is as a tool for researchers to learn more about content labeling and misinformation-deterring strategies across many platforms. Some examples are Google's Top Stories feature [39], and news publishers' search results for reputation check [21].

In the past, researchers have audited YouTube to expose which factors, such as watch history and demographics, have the most influence on whether misinformation is recommended to users [13] as well as identifying user radicalization pathways [30]. YouTube has also been audited to understand how filter bubbles form [37] and what it takes to undo them [35]. [17] conducts a crowd-sourced audit to assess YouTube's stated intention to remove conspiratorial content regarding the 2020 US presidential election from recommendations, while [3] audits the YouTube recommendation algorithm to determine if it played a role in radicalizing users about the 2020 US presidential election. Because of the ever-changing nature of social media and misinformation, [34] advocates for continuous automatic audits.

Many audit studies gather queries using automated resources, such as Google Trends or search auto-complete [13] and videos collected by fact-check sites [32]. While these methods are a great way to focus on the high-volume queries of the moment or target only videos with misinformation, they do not capture the variation of search queries formulated by individuals. Additionally, it may not be the case that Google Trends reflects true population behavior [38]. The two studies mentioned previously, [3, 17], are the only studies that we know of that use crowd-sourced data for their audit of YouTube (in the context of the 2020 US presidential election). We aim to fill this gap in the literature, by crowd-sourcing queries about conspiracy topics, which we then compare against queries and topics gathered without the direct input of search engine users.

2.3 Content labeling and nudges

Platforms like YouTube, Facebook, Twitter, and Instagram have implemented several forms of content moderation to fight problematic content on their respective platforms. One form of content moderation is content labeling, essentially information attached to online posts and websites that is intended to help individuals understand the intent and credibility of the content. In response to the rise in content moderation strategies across platforms, researchers have begun detailing the different forms of content labeling and the implications for both platforms and users [20, 23]. Prior research has analyzed the effect of fact-checking on users from a psychological perspective [24], investigated political bias in content moderation [16], detailed the fact-checking of news publishers via Reviewed Claims on Google search results pages [21], and considered the effect of source-level labeling of political bias on assessing credibility [28]. [33] provides an overview of the strategies used by Facebook and YouTube to prevent the spread of COVID-19 misinformation, including a brief mention of IPs. More recent work has evaluated democratic forms of content labeling that are not centralized to companies, including browser extensions with crowd-sourced assessments of credibility [15] and citations [11]. The latter study includes a survey of YouTube users on which features of a video they rely on to determine credibility.

Similarly to our auditing work, [18] studied IPs in the context of their effect on the viewership of anti-vaccination content on YouTube. [32] investigates the frequency of IPs in relation to queries for watch page IPs, the links for which were selected for containing misinformation as rated by a fact-checking site. To the best of our knowledge, [32] and [18] are the only other works to date that have focused on YouTube's IPs. While [18] investigates the effect of IPs on viewership and [32] conducts a brief study on watch page IPs, our work explores the extent to which SERP IPs are implemented on YouTube by using a broad range of conspiratorial themes as well as understanding the consistency, extent, and coverage of IPs.

3 DATA AND METHODS

As described in the literature review, other researchers who have audited YouTube have carefully curated the list of queries to input, to cover certain desired topics (e.g., most popular conspiracy theories). Given the dual purposes of our study (understanding IPs but also experimenting with the auditing instrument), we decided to take a human-centered perspective and try to capture YouTube's response to queries that different individuals (and not the researchers) would type in the search box. Our goal was to gather variations of query formulations about conspiracy theories, to find out how sensitive YouTube's algorithms are to such variation. Thus, our audit methodology is composed of two parts: 1) crowd-sourcing the query phrases from Amazon Mechanical Turk (AMT) workers, and 2) performing searches with these queries in a variety of locations over multiple weeks.

3.1 Crowd-sourcing query phrases

In the following, we give details about the survey that we used to gather query phrases from the crowd.

3.1.1 Survey Composition. Our survey, which was composed on Qualtrics, had two questions and a demographics section. The first question asked respondents whether they had ever searched about conspiracy theory topics. 219 participants answered "Yes" and 90 survey participants responded with "No". We asked this question to get a sense of how "knowledgeable" people are about the topic. The second question provided them with the following prompt: "Please think of any conspiracy theory topics, then enter below three phrases that you have used or would use to search for those topics." There were no additional instructions to report three different subjects or queries related to specific topics. This was intentional to ensure some ecological validity, by not putting any additional constraints.

After this question, we collected demographic information such as political affiliation, education, age, gender, and the self-reported frequency with which the participant received their news from YouTube. We collected this information in order to ensure that our sample is somewhat representative of the general YouTube audience and that the results capture diversity across various dimensions. In this paper, we do not use this demographic information for any other purposes.

3.1.2 Survey Distribution and Results. The survey was distributed using AMT after gaining approval from our institution's IRB (Protocol #22092R-E, effective December 20, 2021). Subjects were limited

to residing in the United States when taking the survey, having a human intelligence task (HIT) approval rate of over 85 percent, having 50 or more approved HITs, and not having taken the survey before. The workers were compensated $0.50 for their participation. A total of 325 respondents took the survey between December 2021 and March 2022. After the removal of duplicate responses, we were left with 309 participants. From these participants, we collected 927 queries. We considered some responses as nonsense and therefore removed them from our analysis. Removing repeated queries and nonsense queries resulted in 620 unique queries. Adding together the number of times each unique query repeated in the dataset amounts to 892 total queries, indicating some consensus in popular conspiracy topics.

The AMT survey participants described themselves politically as 50% on the left, 27% on the right, and 22% moderate. 43% have bachelor's degrees and 42% are between the ages of 30-39. Gender was roughly equal, with 52% of respondents identifying as men. Of our survey participants, 30% percent reported using YouTube for news frequently, and 65% percent that they did so infrequently.

3.2 Auditing YouTube

After collecting the search phrases through our survey, we pre-processed them by lowercasing and removing the '/' character.[4] We identified 620 unique search phrases, which then were used as inputs for our audit. We aimed to study the extent of IPs across a variety of topics produced by real people without external influence like the autosuggestion feature in the search bar. Such a feature may not be used extensively by YouTube searchers interested in conspiracy theories given that YouTube is likely to filter out problematic queries. Additionally, this method of crowd-sourcing queries for a controlled audit protects participants from research methodology that may be intrusive, such as harvesting real-time user data.

To study the effect of geolocation (that is, whether computers located in different geographical areas in the United States will see the same results or not), we set up virtual machine instances on Google Cloud representing multiple geographical locations in the United States.[5] Using Google Cloud servers is a method used by other YouTube audit studies as well, such as [13]. To make sure that the time (morning, afternoon, evening) did not affect the results, the audits were conducted synchronously in all locations. That is, the script that searched YouTube started at the same time and the queries were in the same order in all machines.

Then, to study whether the presence of IPs on search pages changes over time, we repeated our data collection longitudinally on 16 different dates, roughly weekly between March to July 2023.

The setup of the audit was the following: the same Python script and set of queries were stored in each virtual machine instance. On a given date, all instances started running the script at the same exact time. The script created a new instance of the Chrome browser in incognito mode, controlled by Selenium, which then opened YouTube on the browser, searched for the query phrase, stored the SERP as an HTML page, and then closed the browser. This process was repeated for each query. We want to make it explicit that this audit did not use any Google accounts, so the results are not personalized to any particular account. While the lack of personalization might affect how YouTube algorithms work, we did not want to add more variability to our auditing setup. Our goal was to only focus on location, time, and query.

For three locations, some dates did not collect all 620 expected SERPs on all dates. These three locations were removed from the data completely to ensure all locations collected the same amount of data for all dates. Before removal, we had 161,139 SERPs. We were left with 138,880 SERPs from 14 locations on 16 dates using 620 queries that were used for the analysis within this paper.

After the completion of the data collection, the HTML pages of SERPs were scraped to check for the presence of SERP IPs and collect the content of the IP.

3.3 Grouping queries

For the purposes of understanding the kinds of conspiracy theories that YouTube is prioritizing for content labeling, the unique queries gathered by the survey were grouped in two ways: using Wikipedia-based categorization and handpicked topics (i.e. manually formulated). We analyze the proportion of Wikipedia categories and handpicked topics with IPs. Further, we use the groupings to determine how YouTube covers topics that are popular among the survey participants using context IPs.

Table 1: Handpicked topics with examples of queries grouped under that topic. They show the variation in query formulation by our survey participants.

Handpicked topic	Example queries
911	september 11th conspiracy, 911 truth, 911 inside job, spare change, who did 911, 911 is fake, jetfuel cant melt steel beams, building 7 twin towers, 911 was an inside job
election	2020 election scandal, trump election, votes faked, 2020 vote, voting scams, trump won 2021, biden lost, stolen election
covid	covid not real, covid fake, remvesidir, ivermectin use for covid, covid19, wuhan lab leak theory, covid 19 origins, covid hoax

We determined the **Wikipedia categories** by using the Wikipedia article titled "List of conspiracy theories,"[6] which outlines broad categories of conspiracy theories with examples under each category. Then, we manually sorted queries into 11 broad categories using examples given in the article as guidelines for sorting.

All unique queries were grouped again into 92 topics handpicked by the first author based on the queries in the dataset. We will refer to this grouping as **handpicked topics**. Examples of handpicked topics with queries are shown in Table 1.

[4]Since we save the results of each search as an HTML file with the query name, the presence of the '/' character is problematic, since it affects file path names.

[5]Google Cloud has physical servers in different areas around the world, see here: https://cloud.google.com/compute/docs/regions-zones. The areas used in the audit were as follows: iowa-zone-c, las-vegas-zone-b, las-vegas-zone-c, los-angeles-zone-a, los-angeles-zone-b, northern-virginia-zone-b, northern-virginia-zone-c, oregon-zone-a, oregon-zone-c, salt-lake-city-zone-a, salt-lake-city-zone-b, south-carolina-zone-b, south-carolina-zone-c, south-carolina-zone-d.

[6]https://en.wikipedia.org/wiki/List_of_conspiracy_theories, Accessed 12/6/23.

Table 2: Left subtable: Top 10 queries (out of 620 unique) generated by crowdworkers with the number of times they were repeated. Middle subtable: Top 10 handpicked topics (out of 92) and the total number of queries in each group. Left subtable: Wikipedia categories and the total number of queries in each group. For all query groupings refer to the GitHub repository for this project, linked in Section A.

Query	Count
flat earth	22
illuminati	19
aliens	17
qanon	13
jfk	12
jfk assassination	12
moon landing	12
fake moon landing	12
new world order	8
moon landing fake	8

Handpicked topics	#queries
aliens	98
jfk assassination	88
moon landing	66
global elite	55
911	53
covid	52
election	39
us government	34
vaccine	33
flat earth	29

Wikipedia categories	#queries
government politics and conflict	221
outer space	164
deaths and disappearances	124
science and technology	108
medicine	88
economics and society	86
ethnicity race and religion	9
sports	3
business and industry	3
aviation	2

We acknowledge that these groupings are subjective and may differ depending on how a conspiracy theory is interpreted. For example, *area 51* was assigned the Wikipedia category **outer space** but could also fit under **government, politics, and conflict** considering the theory revolves around both aliens and the US government's involvement. To make our groupings more objective, queries were categorized by multiple researchers, and any discrepancies were resolved through conversations.

To provide an aggregated overview of the queries and the topics, we show them grouped in three ways in Table 2. In the left subtable, we show the top 10 of the unique query phrases, sorted by the number of times they were repeated in the survey response data. In the middle subtable, we show our most popular handpicked topics sorted by the number of queries included in that topic, and in the final subtable, we do the same for the Wikipedia categories. These groups help to give a sense of what conspiracy theories are most popular in our sample, a proxy of the American public in general.

4 RESULTS

4.1 RQ1: How prevalent are IPs on YouTube?

We analyzed 138,880 SERPs collected using 620 input queries across 14 locations on 16 different dates. Of these, 23,100 SERPs (16.63% of the dataset) had an IP present. 114 queries triggered a total of 49 unique IPs. On average, 103 out of 620 SERPs showed IPs on any given day at any given location (SD = 1.89).

Three different types of IPs were found in our dataset: context (21,162 SERPs, 28 unique), fact-check (1,892 SERPs, 17 unique), and news (46 SERPs; 4 unique). For any given day at any given location, context IPs were the most prevalent type (mean = 95, SD = 0.87). The next most prevalent type was the fact-check IP (mean = 8, SD = 1.16). News IPs were the least prevalent (mean = 0.2, SD = 0.57).

Given that context IPs are intended to cover a range of queries within a topic (while fact-check and news IPs are invoked in response to specific statements or queries searching for news), we were expecting that the context IPs that are shown in the dataset will cover all queries that are relevant to the IP. For handpicked topics containing at least one query invoking a context IP, there are 62,944 collected SERPs expected to have an IP. The actual number of SERPs with a context IP (21,162) covers 33.6% of what we expected.

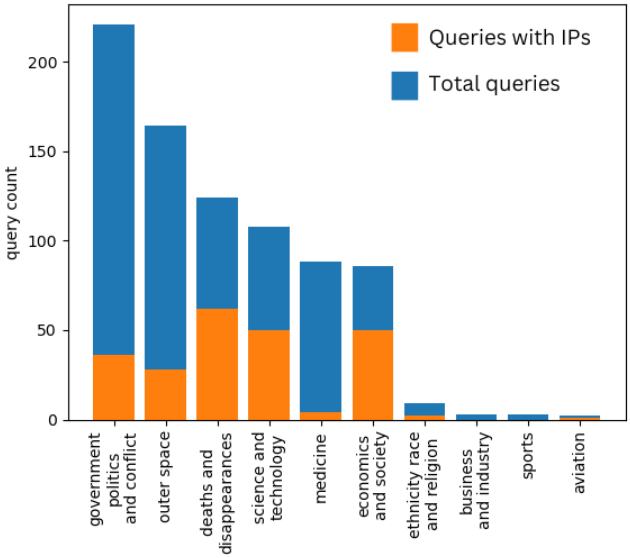

Figure 3: Proportion of all Wikipedia categories with any type of IP. Blue represents the total number of queries, orange represents the number of queries that invoked an IP at least once. The mean number of IP-invoking queries across categories is 23 with a standard deviation of 23.

Additionally, out of the 114 queries that showed IPs, 22 queries showed IPs only partially (i.e., showed up for at least one day or location, but not the rest). We go more in-depth about these queries in Section 4.2. Because YouTube has shown an IP for these queries at least once, this indicates that it recognizes these queries as related to problematic content. For these 22 queries, we calculated the number of times the IP did not occur over the 16 dates and 14 locations: 2,436. This corresponds to a 49.4% absence rate for queries that are recognized as problematic.

Table 3: Examples of context information panels and some of the queries that invoked the IP at least once along with similar queries (in syntax and/or semantics) that did not invoke any IPs.

Context IP	Queries that invoke the IP	Queries that don't invoke any IP
Britannica, *assassination of John F. Kennedy*	jfk assignation, jfk assassination, Kennedy?, jfk shooting, jfk assassination truth, kennedy death	jfk two shooters, who shot john f. john f kennedy, jfk was killed by multiple people
Wikipedia, *QAnon*	q anon, qanon, what is q-anon?, qanon and the election, qanon beliefs	q drops, pedophile ring, q, origin of q
Britannica, *Area 51*	area 51, area 51 aliens, aliens area 51, conspiracy area 51, what is in area 51	area 41, area 52, area 54
Britannica, *illuminati*	illuminati runs the world, illuminati, illumninati, illuminati conspiracy theory, bavarian illuminati	luminate, illuminate, secret society plans, secret organization

Further, context IPs in particular would be expected to cover popular topics prone to conspiracy theories (i.e. topics for which our participants formulated many queries). Four of the top 10 hand-picked topics (**911**, **covid**, **election**, **vaccine**) did not have context IPs associated with them (see example queries for some of these topics in Table 1).

Similarly, when analyzing the prevalence of all IPs as the proportion of queries in Wikipedia categories, see Figure 3, we find that the number of queries with IPs within Wikipedia categories is not related to how popular the category is, as determined by the total number of queries in the category. Concretely, Wikipedia categories have an average of 23 invoking queries (SD = 23). However, the majority of queries for the two main categories: *government, politics, and conflict* and *outer space* don't appear to have triggered any IP.

To summarize the findings of this section, we can say that in the context of our audit, IPs do not appear to be very prevalent on YouTube among queries formulated with the intent of finding conspiratorial content. This claim is supported by the fact that a relatively small percentage of our dataset of SERPs displayed IPs, IP coverage did not follow patterns of query and topic popularity, and even queries that were marked as related to problematic content by YouTube's system did not show IPs across all dates and locations.

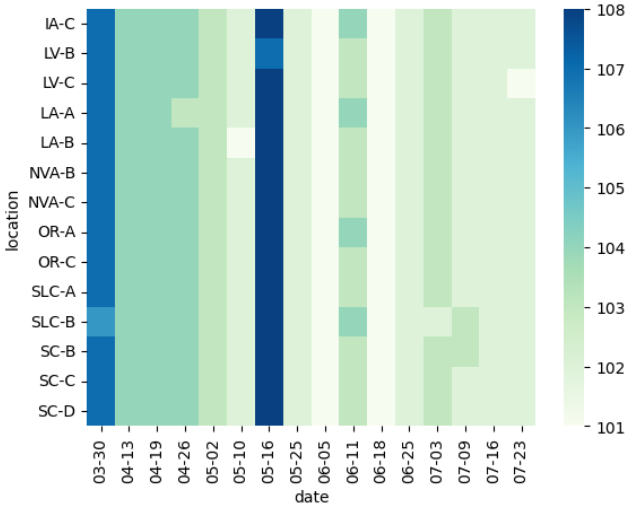

Figure 4: Information panel (IP) distribution across dates and locations. Colorbar values represent the number of SERPs with an IP present from 620 input queries. X-axis values are the dates of data collection in 2023 from March to July.

4.2 RQ2: How do IPs change across time and location?

This investigation examines the effect of date and location on IP occurrence. We plotted the number of IPs for each round of data collection and observed that the highest occurrence of SERPs with IPs on one day in one location was 108 and the lowest was 101 (Figure 4). The date of data collection had more influence on the number of SERPs with a present IP than location, though there was a slight variation in the number of IPs across locations within a small subset of dates. The heatmap in Figure 4 makes this clear: there is little variation vertically (each column is mostly mono-chromatic - meaning the IP number was constant across all locations), meanwhile, there is large variation horizontall, showing that the number of IPs changed from one date to the next.

Because the number of queries with an IP on any given date and location varies, this means that a subset of queries are invoking IPs inconsistently. We became interested in investigating if there is a pattern associated with these inconsistencies. Figure 5 depicts all 22 queries that inconsistently showed IPs across dates. Eight of these 22 queries are inconsistent across both date and location. Qualitatively, there is no clear pattern as to why these queries are inconsistently showing IPs.

Additionally, the date influenced the type of IP shown. The vast majority of queries (112) showed consistent IP types across all SERPs collected using that query. For example, the IP associated with *jfk assassination* was always a context IP. However, 2 queries did not follow this pattern and invoked either a fact-check or context IP depending on the day. Meanwhile, the query *chemtrail evidence* invoked a fact-check on 6/18/23, but invoked a context IP on every other day of data collection. The query *what are chem trails?* invoked a fact-check IP on 9 out of 16 dates and a context IP on 7 out of 16 dates, not clustered together.

Thus, date is more influential on IP occurrence than location, though there was still very slight variation between locations for

Table 4: Examples of queries that invoked fact-check information panels.

Query	Fact-check IP
covid vaccine nanobots	Politifact, *'Transhumanism nanotechnology' COVID-19 vaccine conspiracy theory is Pants on Fire*
is epstein alive?	Politifact, *Jeffrey Epstein is still dead*
melania trump double	USA Today, *Fact check: Images show Melania Trump, not a body double*
celebrity clones	Politifact, *Any resemblance is purely coincidental: These pictures aren't proof of celebrity clones*
covid not real	USA Today, *Fact check: The COVID-19 pandemic is not a hoax*

some queries. About 1/5 of all queries with an IP varied by date and less than 1/10 varied by location. This suggests that in future audits it might not be necessary to use a complex setup to account for different locations as that is less likely to influence the results. Meanwhile, periodically collecting data over time will be very important to capture changes in coverage.

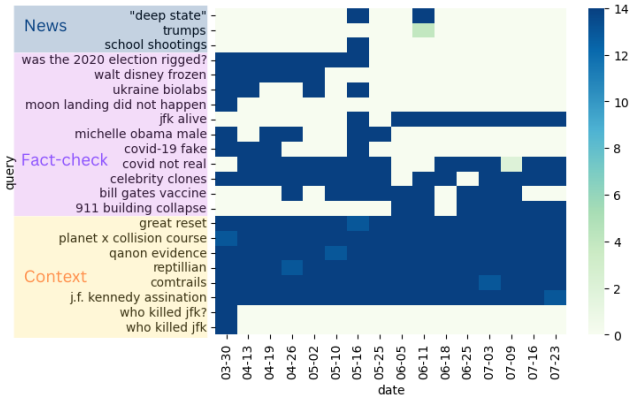

Figure 5: Information panel distribution across date for queries that invoke information panels for only some dates. Colorbar values represent the number of locations with information panels out of 14 total locations. X-axis values are the dates of data collection in 2023 from March to July.

4.3 RQ3: What is the relationship between queries and IP types?

Three types of IPs appeared in the dataset: context, fact-check, and news. Each type has a different purpose, as stated by YouTube[7]. We broke down the data by IP type and conducted a qualitative analysis of the queries that invoke IPs for the purpose of investigating what about a query triggers an IP.

4.3.1 Context IPs. Context IPs made up the majority of all IPs in the dataset, with 21,162 SERPs having context IPs or 15.24% of the dataset. We have a total of 28 unique context IPs triggered by 97 queries. IPs tend to cover more than one query (Table 3).

[7]https://support.google.com/youtube/topic/9257092?hl=en&ref_topic=9257408&sjid=3128292846590594282-NA, Accessed 12/6/23.

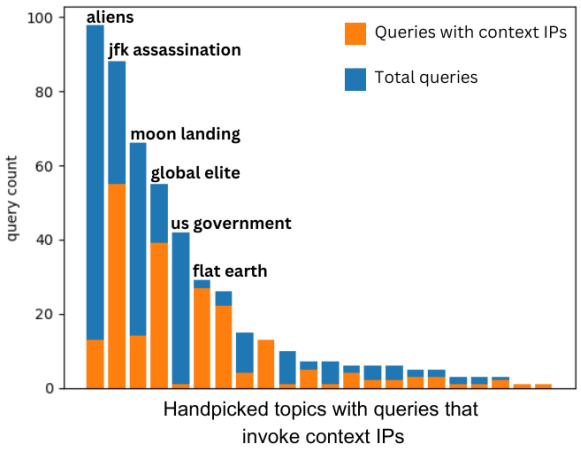

Figure 6: Proportion of queries in handpicked topics that invoke a context information panel at least once. Handpicked topics with no queries that invoke a context IP are not shown. Orange represents the number of context-IP-invoking queries. Blue represents the total number of queries in a topic. The mean number of queries that invoke context IPs is 10 with a standard deviation of 14.

On average, one context IP is invoked by about 4 queries with a standard deviation of 6. The number of queries that invoke one context IP ranges from 1 to 30.

30 of these queries invoked one IP to the Encyclopedia Britannica page "assassination of John F. Kennedy" at least once in the dataset (this is the same IP shown in Figure 1). However, there are 50 queries within the handpicked topic **jfk assassination**. When observing the 30 queries that invoked a context IP, all included John F. Kennedy's name or the acronym of his name, which is also present in the text of the IP. Two of the 20 queries within the handpicked topic **jfk assassination** that did not invoke a context IP do not include the name of John F. Kennedy: *magic bullet theory* and *the magic bullet theory*. However, 18 of the 20 were qualitatively very similar, with some reference to the event. For example, *who shot john f. kennedy?*, *government killed jfk*, and *john f kennedy inside job*.

Table 5: All queries that invoke news information panels.

Query	News IP
"deep state"	The Daily Best, *Trump Rants About 'SCUM' and 'COCKROACHES' After Durham Report*
	Ledger-Enquirer, *Donald Trump's visit to Columbus for Georgia GOP convention sparks protest, celebration*
	Washington Post, *How Garland's release of Trump-Russia probe report differed from Barr's*
school shootings	CNN, *Gunman who killed 3 people and injured 6 in Farmington, New Mexico, was 18 years old and used three firearms, police say*
trumps	Ledger-Enquirer, *Donald Trump's visit to Columbus for Georgia GOP convention sparks protest, celebration*

A similar behavior is true for all queries that invoked a context IP, wherein keywords in the query match part of the IP's text, except for the query *global warming*, which invokes an IP that does not mention the phrase directly in the text or in the article but includes the synonym phrase *climate change*.

To determine the relationship between the popularity of handpicked topics and how often the queries within those topics invoke IPs at least once, we compared the total number of queries in a topic to the number of queries that invoked IPs (Figure 6). We expected to see a functional relationship between the number of IP-invoking queries and the total number of queries in the handpicked topics, but the proportion varied.

4.3.2 Fact-Check IPs. Fact-check IPs made up the next largest share of IPs in the dataset, with 1,892 SERPs having a fact-check IP or 1.36% of the dataset. Compared to context IPs, the number of fact-check IPs roughly matched the number of invoking queries. 16 queries invoked a total of 17 fact-check IP URLs. Examples of these queries are shown in Table 4.

Similarly to the context IP queries, the queries that invoked fact-checks had at least one keyword that matched the IP text. However, we found that queries invoking fact-check IPs tend to have more than one keyword in the query phrase matching the text of the IP.

Only three queries invoke a fact-check IP on every day of data collection in every location: *covid vaccine nanobot*, *is epstein alive?*, *melania trump double*, indicating that fact-check IPs are less reliable than context IPs.

Two queries invoked more than one fact-check IP URL. The first query *michelle obama male* invoked three different IP URLs that change according to the date. At the start of the data collection period in March of 2023, the query resulted in a SERP with an IP that linked to a Politifact.com article published in May 2021. On 4/19/23, the IP linked to a different Politifact.com article published February 2023. The IP changed again 5/16/23 to a Politifact.com article published just a few days before data collection, 5/8/23. Each of the three articles debunks a different piece of evidence for the conspiracy theory.

The second query *jfk alive* links to two different fact-check articles, both from Politifact.com with similar content, published on the same day. One article shows up for 3 days of data collection in the month of June, while the other shows up for 6 scattered days across the data collection period.

4.3.3 News IPs. News IPs were the most infrequent type of IP in the dataset, with only 46 SERPs having a news IP or 0.03% of the dataset. Similar to fact-check IPs, the number of news IPs roughly matched the number of invoking queries. 3 queries invoked a total of 4 news IPs. All three queries and the IPs they invoke are shown in Table 5.

The news IPs were highly dependent on time, with IPs showing up for one day for two queries (*trumps* and *school shootings*) and two days for the third query (*"deep state"*[8]).

The first query *school shootings* invoked an IP on 5/16/23 linked to a CNN.com article reporting on a school shooting that occurred on 5/15/23.

The second query *"deep state"* invoked three different news IPs. On one day, one location showed an IP that linked to a WashingtonPost.com article and 15 locations showed an IP that linked to a TheDailyBeast.com article. On the second day the IP was present, 6/11/23, the same query showed an IP linked to a Ledger-Enquirer.com article. On the same day, the third query *trumps* invoked the same Ledger-Enquirer.com IP. Each of these three articles reported on completely different events relating to Donald Trump. None of the articles included the phrase "deep state".

Interestingly, a separate query *deep state* was present in the list of queries (without quotation marks) and did not invoke an IP across the data collection dates and locations.

These qualitative case studies reveal that related keywords are seemingly required for the presence of context and fact-check IPs, but this relationship is less obvious for news IPs. While keywords might be required, they are not necessarily sufficient. Some queries with high word overlap did not show IPs. Additionally, some queries both had word overlap and showed IPs on some days, but not on other days. More research is needed to understand the variables that connect queries and IPs.

A clear result from this work is the variation between IP type and date. When a query is connected to a context IP, it will show in almost all conditions. However, fact-check IPs are much more dependent on date, though the pattern of this relationship is not obvious. News IPs are highly dependent on date. This relationship is

[8]The query included quotation marks in the dataset and was inputted to YouTube in this form.

somewhat clear and intuitive based on the purpose of the news IPs: the news story within the IP is published and shown to users during the time frame of the breaking news event. In one example, the query *school shooting* matched an IP giving information on a school shooting that happened the day prior to data collection. However, this is just one example out of only 3 queries that invoked IPs. More news IPs need to be analyzed to understand their function and reach. This is why in the following section, we discuss an extension of the audit methodology intended to inform future audits that will try to replicate and expand upon these findings.

5 DISCUSSION, LIMITATIONS, FUTURE WORK

5.1 Towards News-Cycle-Aware Audits

One of our goals for this study was to investigate how to best conduct meaningful audits of YouTube, that allow us to capture its changing behavior toward problematic content, in a quasi-realistic way. We experimented with three variables: crowd-sourced queries, time, and location (the geographical location of machines on which the audit was carried out). We find that IP presence varies more over time than location. Some of this variation is due to links updating, particularly for news and fact-check IPs. Other times, an IP disappears completely with no indication as to why.

We further tested this time-dependency hypothesis by collecting data for an additional two dates (10/30/23 and 11/07/23) across the same 14 locations using the same 620 input queries. This round of data collection resulted in 17,360 SERPs and will be referred to as the "2-week follow-up dataset".

Comparing the 2-week follow-up dataset to the original 16-week dataset reveals a jump between 7/23/23 and 10/30/23 in the highest number of IP-invoking queries in one day from 108 to 130. In total, 32 queries invoked IPs in the 2-week follow-up dataset that did not invoke any IPs in the 16-week dataset. Most of these added queries (26) were related to the 9/11 topic, while the other 6 were unrelated to each other. One new context IP was added (a Britannica article on 9/11 attacks) along with 4 fact-check IPs (on various topics). No news IPs were encountered. At the same time, 17 total IPs had "disappeared", including all 4 news IPs and 13 fact-check IPs from the 16-week dataset.

Thinking about these results in the context of global news events, we wonder if the addition of the IP showing the Encyclopedia Britannica article on the 9/11 attacks was related to the attack on Israel by Hamas on 10/7/23, just a few weeks before we collected our follow-up data. We think this is plausible considering the comparisons made between the 9/11 attacks and the events in Israel [19].

Our findings indicate that news IPs are highly dependent on time. This matches YouTube's description of news IPs that they show up for "developing news stories"[9]. Fact-check IPs display a similar dependence on time, with articles updating or disappearing altogether for some dates.

Given the dependence on time for IPs overall and our hypothesis that YouTube may be updating IPs in response to real-world events, based on our follow-up data, we believe that it is important for researchers to develop methods that enable auditing that is acute and responsive to real-world events. We refer to this as "news-cycle-aware auditing". This type of auditing would allow researchers to capture changes in content labeling that occur in response to problematic content arising on platforms as a result of real-world events. A challenge for such audits is the choice of keywords. Typically, one cannot predict ahead of time what queries will be used by YouTube users to search for a developing story, but also, what old conspiracy theories might resurface due to perceived similarities to an unfolding event. This is why more research is needed to design methods for news-cycle-aware audits.

5.2 Problematic content and content labeling

Each type of IP (context, fact-check, news) seems to rely on different occurrence criteria, with one common thread: keywords.

We observed that queries invoking fact-check and news IPs typically matched multiple keywords in the article being shown, while context IPs had at least one keyword in the query phrase. This is also indicated by the fact that one context IP typically covers multiple query phrases while the fact-check and news IPs are targeted to specific query formulations. This matches YouTube's explanation of showing fact-check IPs in response to "specific claim[s]"[10].

While keywords seem to be a requirement for YouTube's algorithm to recognize the need for an IP, it does not always guarantee the presence of an IP. We found that some queries that invoke IPs are qualitatively similar to queries that do not invoke any IP. This inconsistent triggering of IPs is puzzling and is one of the areas that needs to be researched further.

Once a query has triggered an IP, a process of source selection, article selection, and panel text selection needs to be completed before an IP is displayed on the SERP. The methods YouTube uses to complete these processes are also unclear based on the results of this audit.

For example, YouTube states that the sources for news IPs are chosen using a range of criteria, including local relevance, relevance to the topic, date posted, reporting intent, association with websites whose content surfaces on Google News, and more. In our data, we encountered four sources for news IPs: WashingtonPost.com & CNN.com (widely recognized news sources worldwide) and TheDailyBeast.com & Ledger-Enquirer.com (low levels of recognition and disputed factual accuracy). Given the importance of accurate information during developing news events, the choice of the news source to show in the IP is very important.

Further, many queries in the dataset seem to seek information about a specific popular claim, like *earth is flat*, but did not invoke a fact-check IP. YouTube gives this explanation[11]:

> If a fact-check doesn't appear, it may be because an eligible publisher hasn't published a fact-check article relevant to your search.

However, a quick search on Politifact.com for the phrase "earth is flat" reveals many articles on the topic.[12]

We collected the crowd-sourced queries by explicitly asking crowdworkers to provide a query they would use to search for a

[9] https://support.google.com/youtube/answer/9057101?hl=en&ref_topic=9257092&sjid=3128292846590594282-NA, Accessed 12/6/23.

[10] https://support.google.com/youtube/answer/9229632?hl=en&ref_topic=9257092&sjid=3128292846590594282-NA

[11] https://support.google.com/youtube/answer/9229632?hl=en&ref_topic=9257092&sjid=3128292846590594282-NA, Accessed 12/6/23.

[12] https://www.politifact.com/search/?q=earth+is+flat, Accessed 12/6/23.

conspiracy theory. In some cases, the query phrase does not indicate that intent explicitely in its formulation (for example, *jfk* or *moon landing*). As researchers, we know the intent of the query, but the algorithm does not recognize the connection to problematic content and therefore does not show an IP. In these cases, we would say that the algorithm is not sensitive enough to queries with problematic intent, either by design or not.

We speculate that the YouTube algorithm may have some threshold for recognizing queries with conspiratorial intent that is higher than what we would expect. We hypothesize that this is the case because YouTube might not want to expose the YouTube audience to conspiracy theories. That is, YouTube might be using an editorial policy known as "strategic silence" [4]. Testing such a hypothesis might require a different audit setup, especially one that would take into account a user's prior search behavior.

In addition to calling for the YouTube algorithm to add content labels more consistently and with better coverage, we believe the content of the IPs should be meaningful. Context IPs themselves often don't warn the user that the topic may be connected to problematic content. Of the 31 context IPs, only 7 of them explicitly mention the topic's connection to conspiracy theories. And only a single context IP uses the word "disproven". Thus, it's unclear how effective context IPs are at deterring users from seeking and believing problematic content. Future work should evaluate the efficacy of different IP types in deterring YouTube users from uncritically consuming or believing problematic content.

Given the importance of YouTube on the information environment and the potential for problematic content to have real-world harm (i.e., the Marjory Stoneman Douglas shooting "crisis actors" narrative), informing consumers about the truth consensus with respect to certain topics is incredibly important. YouTube seemingly recognizes the need for a system that supports users in understanding the content on the platform, thus implementing and updating IPs over the years. In our view, if YouTube is serious about the value of the IP strategy in discouraging the consumption of problematic content within the platform, the implementation of the policy should be clearly stated or, at the very least, easily discerned from interacting with the platform. We provide evidence in this paper that the criteria YouTube uses to decide if and when an IP is shown to users are obscure.

5.3 Limitations

While our goal with crowd-sourcing queries was to capture the variation in query formulation by real people searching YouTube for conspiracy theories, the queries were ultimately prompted by our survey, as opposed to having been extracted directly from a user's interaction with YouTube, for example, through a browser extension. Furthermore, this process also does not capture a user's "on-the-fly information need" which is usually prompted by external factors like coming across an unknown (conspiracy) theory in everyday life. Despite these shortcomings related to the ecological validity of the query selection, our queries cover a wide range of topics and formulations not addressed in other studies. Further, we believe that this strategy has some unique benefits, including avoiding collection of user data and thereby protecting participant privacy. In this study, we were interested in focusing on a few key variables (time, space, and query) in order to understand the general implementation of IPs. IPs as a content labelling strategy are intended to warn users coming across potentially problematic content, so they should be available regardless of individual user demographics or behavior, an additional reason to disregard personalization as a variable in this audit.

Another limitation is related to the lack of news-relevant queries, which led to a very small sample size for news IPs in our dataset. This is due to the time gap between query collection and SERP collection. For one to form a comprehensive study of news IPs on YouTube, a news-cycle-aware audit methodology must be adopted.

Finally, our audit is limited to the United States and the English language. While YouTube has declared that the IP feature will be available in other countries and languages other than English, our audit did not consider other localities. However, our audit methodology (the use of virtual machines in Google Data centers located across the world, can be used to perform audits of YouTube in other countries.

5.4 Future work

Though this study examines only variations in the occurrence of YouTube's IPs, there are many possible avenues for continuing to understand how this intervention is shaping the information landscape on the YouTube platform and broader web.

One important consideration is how the presence of IPs effects a users watch behavior on YouTube. A user-centered study may help to understand real-time video-watching behavior when IPs are present on YouTube's search results page or on video watch pages. Some research has already been done to characterize which features of YouTube videos watchers utilize to assess the credibility, including some interpretation of the IP design [11]. An additional consideration could be how demographic information and account personalization influence conspiracy-related query searches; how often people are exposed to IPs when browsing YouTube; and the perceived value of IPs as credibility signals.

We decided to focus on YouTube due to the immense popularity of the platform and its previous history of hosting conspiratorial content. However, understanding content labeling on YouTube has implications for the broader information landscape. Researchers are working to centralize information about the types of interventions employed online to decrease the spread of misinformation. One example is an online toolbox of interventions, including data about each conceptual intervention like scalability [20]. This study can be used to inform such efforts, which will in turn provide policymakers with knowledge of effective and practical forms of content moderation.

6 CONCLUSION

Conspiracy theories have always thrived on YouTube and one mechanism that YouTube has chosen to warn users about them is through information panels (IPs). We designed a longitudinal audit to study the presence of IPs on YouTube search pages across 14 U.S. locations and 16 weeks using 620 crowd-sourced queries seeking information on a range of conspiracy theories. The audit revealed the presence of three types of IPs: context IP, fact-check IP, and news IP. Our results indicate that IPs are not extensive (covering only 16% of the

dataset), with context IPs being the most prevalent and consistent across time, while fact-check and news IPs vary significantly across time. The latter result demonstrates the need for a new type of audit, a news-cycle-aware audit, to accurately capture time-sensitive content labeling. At the same time, we find that IP presence does not vary by location within a country, suggesting that audits can downplay (or even omit) location as a variable.

This paper finds that YouTube is inconsistent in showing IPs in response to conspiracy-related queries. Such inconsistencies are inadequate, because YouTube has already identified the subject to be potentially problematic, yet it allows some of this content to be shown without providing content labeling. This seems to be related to the hypothesized keyword-matching mechanism for invoking IPs. As argued in the paper, the presence of certain keywords seem to be a requirement for queries to invoke IPs. Context IPs are invoked when a query has at least one keyword, while fact-check and news IPs are tailored to queries with multiple keywords. Although this study sheds light on some of the details of how the IP-selection algorithm might work, further research is needed to better understand its behavior.

ACKNOWLEDGMENTS

We are grateful to all members of Wellesley Cred Lab, in particular to Ropah Shava, Jessi Kim, Audrey Yip, and Fridah Ntika. We also acknowledge funding from NSF grant ISS 1751087.

REFERENCES

[1] Joachim Allgaier. 2016. Science on YouTube: What users find when they search for climate science and climate manipulation. *arXiv:1602.02692 [cs]* (April 2016). http://arxiv.org/abs/1602.02692

[2] Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 74 (apr 2021), 34 pages. https://doi.org/10.1145/3449148

[3] James Bisbee, Megan Brown, Angela Lai, Richard Bonneau, Joshua A Tucker, and Jonathan Nagler. 2022. Election Fraud, YouTube, and Public Perception of the Legitimacy of President Biden. *Journal of Online Trust and Safety* 1, 3 (Aug. 2022). https://doi.org/10.54501/jots.v1i3.60

[4] Joan Donovan and Danah Boyd. 2021. Stop the presses? Moving from strategic silence to strategic amplification in a networked media ecosystem. *American Behavioral Scientist* 65, 2 (2021), 333–350.

[5] Marc Dupuis, Kelly Chhor, and Nhu Ly. 2021. Misinformation and Disinformation in the Era of COVID-19: The Role of Primary Information Sources and the Development of Attitudes Toward Vaccination. In *Proceedings of the 22nd Annual Conference on Information Technology Education* (SnowBird, UT, USA) *(SIGITE '21)*. Association for Computing Machinery, New York, NY, USA, 105–110. https://doi.org/10.1145/3450329.3476866

[6] Marc Faddoul, Guillaume Chaslot, and Hany Farid. 2020. A Longitudinal Analysis of YouTube's Promotion of Conspiracy Videos. *arXiv:2003.03318 [cs]* (March 2020). http://arxiv.org/abs/2003.03318

[7] Human Health Services. 2021. U.S. Surgeon General Issues Advisory During COVID-19 Vaccination Push Warning American Public About Threat of Health Misinformation. *HHS* (July 2021). https://www.hhs.gov/about/news/2021/07/15/us-surgeon-general-issues-advisory-during-covid-19-vaccination-push-warning-american.html

[8] John Herrman. 2018. The Making of a No. 1 YouTube Conspiracy Video After the Parkland Tragedy. *The New York Times* (Feb. 2018). https://www.nytimes.com/2018/02/21/business/media/youtube-conspiracy-video-parkland.html

[9] Amanda Hess. 2019. They Kinda Want to Believe Apollo 11 Was Maybe a Hoax. *The New York Times* (July 2019). https://www.nytimes.com/2019/07/01/science/moon-landing-hoax-conspiracy-theory.html

[10] Andrew Higgins. 2021. The Art of the Lie? The Bigger the Better. *The New York Times* (Feb. 2021). https://www.nytimes.com/2021/01/10/world/europe/trump-truth-lies-power.html

[11] Emelia May Hughes, Renee Wang, Prerna Juneja, Tony W Li, Tanushree Mitra, and Amy X. Zhang. 2024. Viblio: Introducing Credibility Signals and Citations to Video-Sharing Platforms. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 807, 20 pages. https://doi.org/10.1145/3613904.3642490

[12] Muhammad Nihal Hussain, Serpil Tokdemir, Nitin Agarwal, and Samer Al-Khateeb. 2018. Analyzing Disinformation and Crowd Manipulation Tactics on YouTube. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 1092–1095. https://doi.org/10.1109/ASONAM.2018.8508766

[13] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. 2020. Measuring misinformation in video search platforms: An audit study on YouTube. *Proc. of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–27.

[14] Caroline Jack. 2017. Lexicon of lies: Terms for problematic information. (2017).

[15] Farnaz Jahanbakhsh and David R Karger. 2024. A Browser Extension for in-place Signaling and Assessment of Misinformation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 946, 21 pages. https://doi.org/10.1145/3613904.3642473

[16] Shan Jiang, Ronald E. Robertson, and Christo Wilson. 2020. Reasoning about Political Bias in Content Moderation. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 09 (April 2020), 13669–13672. https://doi.org/10.1609/aaai.v34i09.7117

[17] Prerna Juneja, Md Momen Bhuiyan, and Tanushree Mitra. 2023. Assessing Enactment of Content Regulation Policies: A Post Hoc Crowd-Sourced Audit of Election Misinformation on YouTube. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 545, 22 pages. https://doi.org/10.1145/3544548.3580846

[18] Sangyeon Kim, Omer F. Yalcin, Samuel E. Bestvater, Kevin Munger, Burt L. Monroe, and Bruce A. Desmarais. 2020. The Effects of an Informational Intervention on Attention to Anti-Vaccination Content on YouTube. *Proceedings of the International AAAI Conference on Web and Social Media* 14 (May 2020), 949–953. https://ojs.aaai.org/index.php/ICWSM/article/view/7364

[19] Peter Knoope. 2023. Comparing Hamas' attack on Israel and 9/11 - A Counterterrorism Perspective. *International Center for Counter-Terrorism* (2023). https://www.icct.nl/publication/comparing-hamas-attack-israel-and-911-counterterrorism-perspective

[20] Anastasia Kozyreva, Philipp Lorenz-Spreen, Stefan M. Herzog, Ullrich K. H. Ecker, Stephan Lewandowsky, Ralph Hertwig, Ayesha Ali, Joe Bak-Coleman, Sarit Barzilai, Melisa Basol, Adam J. Berinsky, Cornelia Betsch, John Cook, Lisa K. Fazio, Michael Geers, Andrew M. Guess, Haifeng Huang, Horacio Larreguy, Rakoen Maertens, Folco Panizza, Gordon Pennycook, David G. Rand, Steve Rathje, Jason Reifler, Philipp Schmid, Mark Smith, Briony Swire-Thompson, Paula Szewach, Sander van der Linden, and Sam Wineburg. 2024. Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour* 8 (2024), 1044–1052.

[21] Emma Lurie and Eni Mustafaraj. 2020. Highly Partisan and Blatantly Wrong: Analyzing News Publishers' Critiques of Google's Reviewed Claims. In *Proc. of the 2nd Trust and Truth Online (TTO'20)*.

[22] Golshan Madraki, Isabella Grasso, Jacqueline M. Otala, Yu Liu, and Jeanna Matthews. 2021. Characterizing and Comparing COVID-19 Misinformation Across Languages, Countries and Platforms. In *Companion Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 213–223. https://doi.org/10.1145/3442442.3452304

[23] Garrett Morrow, Briony Swire-Thompson, Jessica Polny, Matthew Kopec, and John Wihbey. 2020. *The Emerging Science of Content Labeling: Contextualizing Social Media Content Moderation*. SSRN Scholarly Paper ID 3742120. Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.3742120

[24] Elmie Nekmat. 2020. Nudge effect of fact-check alerts: source influence and media skepticism on sharing of news misinformation in social media. *Social Media+ Society* 6, 1 (2020), 2056305119897322.

[25] Adina Nerghes, Peter Kerkhof, and Iina Hellsten. 2018. Early Public Responses to the Zika-Virus on YouTube: Prevalence of and Differences Between Conspiracy Theory and Informational Videos. In *Proceedings of the 10th ACM Conference on Web Science (WebSci '18)*. Association for Computing Machinery, New York, NY, USA, 127–134. https://doi.org/10.1145/3201064.3201086

[26] Gianluca Nogara, Padinjaredath Suresh Vishnuprasad, Felipe Cardoso, Omran Ayoub, Silvia Giordano, and Luca Luceri. 2022. The Disinformation Dozen: An Exploratory Analysis of Covid-19 Disinformation Proliferation on Twitter. In *Proceedings of the 14th ACM Web Science Conference 2022* (Barcelona, Spain) *(WebSci '22)*. Association for Computing Machinery, New York, NY, USA, 348–358. https://doi.org/10.1145/3501247.3531573

[27] John C Paolillo. 2018. The flat earth phenomenon on YouTube. *First Monday* (2018).

[28] Jinkyung Park, Rahul Dev Ellezhuthil, Joseph Isaac, Christoph Mergerson, Lauren Feldman, and Vivek Singh. 2023. Misinformation Detection Algorithms and Fairness across Political Ideologies: The Impact of Article Level Labeling. In *Proceedings of the 15th ACM Web Science Conference 2023* (Austin, TX, USA) *(WebSci '23)*. Association for Computing Machinery, New York, NY, USA, 107–116.

https://doi.org/10.1145/3578503.3583617
[29] Max Read. 2013. Behind the 'Sandy Hook Truther' Conspiracy Video That Five Eight Million People Have Watched in One Week. *Gawker* (Jan. 2013). https://web.archive.org/web/20130116212311/http://gawker.com/5976204/behind-the-sandy-hook-truther-conspiracy-video-that-five-million-people-have-watched-in-one-week
[30] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. 2020. Auditing radicalization pathways on YouTube. In *Proc. of the 2020 Conference on Fairness, Accountability, and Transparency, FAccT'20*. 131–141.
[31] Kevin Roose. 2021. How a viral video bent reality. *The New York Times* (Sept. 2021). https://www.nytimes.com/2021/09/08/technology/loose-change-9-11-video.html
[32] Shaden Shabayek, Héloïse Théro, Dana Almanla, and Emmanuel Vincent. 2022. Monitoring misinformation related interventions by Facebook, Twitter and YouTube: methods and illustration. (May 2022). https://hal.archives-ouvertes.fr/hal-03662191 working paper or preprint.
[33] Eugenia Siapera. 2022. Platform Governance and the "Infodemic". *Javnost - The Public* 29, 2 (2022), 197–214. https://doi.org/10.1080/13183222.2022.2042791 Publisher: Routledge _eprint: https://doi.org/10.1080/13183222.2022.2042791.
[34] Jakub Simko, Matus Tomlein, Branislav Pecher, Robert Moro, Ivan Srba, Elena Stefancova, Andrea Hrckova, Michal Kompan, Juraj Podrouzek, and Maria Bielikova. 2021. Towards Continuous Automatic Audits of Social Media Adaptive Behavior and Its Role in Misinformation Spreading. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (Utrecht, Netherlands) *(UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 411–414. https://doi.org/10.1145/3450614.3463353
[35] Ivan Srba, Robert Moro, Matus Tomlein, Branislav Pecher, Jakub Simko, Elena Stefancova, Michal Kompan, Andrea Hrckova, Juraj Podrouzek, Adrian Gavornik, et al. 2023. Auditing YouTube's recommendation algorithm for misinformation filter bubbles. *ACM Transactions on Recommender Systems* 1, 1 (2023), 1–33.
[36] Dominik A. Stecula and Mark Pickup. 2021. Social Media, Cognitive Reflection, and Conspiracy Beliefs. *Frontiers in Political Science* 3 (2021). https://doi.org/10.3389/fpos.2021.647957
[37] Matus Tomlein, Branislav Pecher, Jakub Simko, Ivan Srba, Robert Moro, Elena Stefancova, Michal Kompan, Andrea Hrckova, Juraj Podrouzek, and Maria Bielikova. 2021. An Audit of Misinformation Filter Bubbles on YouTube: Bubble Bursting and Recent Behavior Changes. In *Proceedings of the 15th ACM Conference on Recommender Systems* (Amsterdam, Netherlands) *(RecSys '21)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3460231.3474241
[38] Ulrich S. Tran, Rita Andel, Thomas Niederkrotenthaler, Benedikt Till, Vladeta Ajdacic-Gross, and Martin Voracek. 2017. Low validity of Google Trends for behavioral forecasting of national suicide rates. *PLOS ONE* 12, 8 (Aug. 2017), e0183149. https://doi.org/10.1371/journal.pone.0183149 Publisher: Public Library of Science.
[39] Daniel Trielli and Nicholas Diakopoulos. 2019. Search as News Curator: The Role of Google in Shaping Attention to News Information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300683

A MATERIALS

GitHub repository containing the materials for this project: https://github.com/lillgodi/YouTube-Information-Panels-Audit-2023