Towards Understanding and Defending Against Algorithmically Curated Misinformation

Prerna Juneja

A dissertation
submitted in the partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2023

Reading Committee:
Tanushree Mitra, Chair
Chirag Shah
Bill Howe

Program Authorized to Offer Degree:
Information School

University of Washington

Abstract

Towards Understanding and Defending Against Algorithmically Curated Misinformation

Prerna Juneja

Chair of the Supervisory Committee:
Tanushree Mitra
Department of Information Science

Search engines and online social media platforms have become important sources of information for users worldwide. Despite their popularity and ubiquitousness, online platforms are not always trustworthy sources of information. The platforms are driven by black box algorithms that optimize for engagement over the credibility of information. There are increasing concerns that online platforms amplify inaccurate information, making it easily accessible via search results and recommendations. In this thesis, I explore the role of online algorithms in promoting misinformation and design defenses against online misinformation by incorporating human-centered insights from stakeholders such as fact-checking organizations and news agencies. My research recognizes the multifaceted nature of online misinformation and explores the algorithmic, policy, fact-checking, and design aspects of the problem through three distinct research threads.

In the first thread of my research, I investigate and audit online platforms such as YouTube and Amazon to understand the role of algorithms driving these platforms in surfacing and amplifying misinformative content to users. Through the audits, I found that performing certain real-world actions on misinformative content (e.g. watching a conspiratorial video on YouTube, or adding a misinformative book to the cart on Amazon) could lead users into problematic echo chambers of misinformation.

Additionally, I identified vulnerable user populations who could be targets for specific misinformative topics on online platforms.

In the second research thread, I explore ways to support the fact-checking process to combat online misinformation. For this work, I interviewed 14 fact-checking organizations and news agencies across four continents to understand their current fact-checking processes, challenges, and needs. This research establishes fact-checking process as a socio-technical phenomenon, revealing the collaborative efforts of various stakeholder groups and technological infrastructure in facilitating effective fact-checking endeavors. It also highlights the technical, policy, and informational barriers to fact-checking and emphasizes the need for systematic changes in civic, informational, and technological contexts to improve the overall quality of fact-checking.

In the final thread of my dissertation research, I collaborated with Pesacheck, Africa's largest indigenous fact-checking organization, to design and develop YouCred—a fact-checking system that enables monitoring of algorithmically driven online platforms for misinformation. To create YouCred, I incorporated insights from previous research threads as well as the expertise and feedback of Pesacheck's fact-checkers throughout the development and design stages. YouCred specifically facilitates misinformation discovery and credibility assessments on the YouTube platform. It automatically generates search queries related to important events and topics of interest to fact-checkers and also offers an intuitive interface for annotating videos for misinformation. Through a nine-month evaluation period at Pesacheck, YouCred demonstrates its practical value and usefulness for fact-checkers, underscoring the importance of ongoing collaboration between fact-checking organizations and technology developers in combating online misinformation.

In conclusion, this thesis adopts a socio-technical approach to understanding and defending against algorithmically curated online misinformation. It also paves the way for future research in designing interventions to counter algorithmic harm and developing socio-technical systems to address the problem of online misinformation.

"There were pages turned with the bridges burned
Everything you lose is a step you take
So make the friendship bracelets
Take the moment and taste it
You've got no reason to be afraid"
— *Taylor Swift*

Dedication

The dedication of this thesis is split three ways: To those who have shown me kindness and provided unwavering support, to TS's songs that have been my constant companion through the good and bad times, and to you dear reader if you find value in any aspect of my work.

Acknowledgments

I find myself humbled and deeply grateful as I reflect upon the journey that has brought me to this moment. Until now, I never fully appreciated the courage it took for me to leave behind a good job, my home, and my country to come to the U.S. for a Ph.D. Spending 5 years for a degree is a long time and a lot of life happens. And truth be told, it hasn't been easy. Like every significant decision I've made, choosing this path made me gain some things and lose some. In the last five years, I survived several personal and professional hardships. Nevertheless, I am truly grateful that the journey is coming to an end on a positive note. I want all my students, friends, mentors, and family to know that your kindness, encouragement, and support are what made this possible.

I first want to highlight the most rewarding aspect of this journey—mentoring students. Most days, I eagerly looked forward to meeting with my students, and our interactions were the highlights of my day. I want to thank David Xie, Louis Leng, Hayoung Jung, Vincent Zhiyuan Zhou, Stephanie L. Zhang, Alice Zhang, Ankita Khera, Benjamin Ye, Lee Polla, and Ethan Yee with all my heart. It has been an absolute honor to know you and to work with you. Your enthusiasm, energy, and creativity breathed life into the projects we collaborated on.

My lab mates made this journey a lot more pleasant and memorable. Shruti, thanks to you, I always had a second home in Blacksburg and Seattle. Momen, it was a pleasure to collaborate with you. Brian, thanks for your willingness to help whenever needed. Kristen, I deeply admire your strength, clarity of mind, and independent spirit. Neelesh and Saloni, your presence infused our lab with much-needed life, energy, and colors. The last year was a lot of fun because of you two. To my friends, I am always grateful for your encouragement. Parul, Shruti Bansal, Arka, and Harry you all are my biggest cheerleaders. I am also grateful to all my therapists. Seeking therapy has emerged as one of life's truest blessings and a very important part of my personal and professional journey.

I also want to thank several researchers who have made a lasting impact on my

academic journey. First, I want to acknowledge Dr. Mitra who introduced me to an incredible field of research that feels like home, where I truly belong. Your discipline and work ethic have been inspiring. Bringing me to Seattle was like offering me a lifeline, and I am sincerely grateful to you for that. I also want to extend my heartfelt gratitude to Dr. Francisco Servant and Dr. Megan Finn, whose teaching styles have been a deep source of inspiration for me. Additionally, I am grateful to my committee members for their guidance and support. I would also like to express my gratitude to Dr. Isabel Zhang and Dr. Alison Renner for their mentorship, as well as Dr. Eni Mustafaraj, and Dr. Kokil Jaidka, for their support, and words of encouragement over the last couple of years.

I also want to acknowledge the huge role art has played these last five years. Especially, you TS. It feels like all these years, you've been filling pages, writing all these songs, narrating all these stories that deeply intertwine with my own life. With your songs, I have smiled, hoped, danced, and experienced a rainbow of beautiful emotions. I must also extend my thanks to the captivating TV shows that provided a perfect escape during the last five years. Schitt's Creek, Atypical, Extraordinary Attorney Woo, Marvelous Mrs. Maisel, and Anne With An E felt like a warm hug. They allowed me to live vicariously through their characters, evoking laughter, and tears, and giving me a lot of amazing moments during my most trying times.

In the end, I want to say that I'm proud of myself. And while my grad school journey is ending, a personal journey of self-acceptance and self-compassion is just beginning.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

> "Google's search algorithm spreads false information with a rightwing bias—Search and autocomplete algorithms prioritize sites with rightwing bias, and far-right groups trick it to boost propaganda and misinformation in search rankings"—The Guardian [3]
>
> "YouTube more likely to recommend election-fraud content to those skeptical of the 2020 election"—The Hill [5]
>
> "YouTube is still suggesting conspiracy videos, hyperpartisan and misogynist videos, pirated videos, and content from hate groups following common news-related searches."—The BuzzFeed [4]
>
> "An Anti-Vaccine Book Tops Amazon's COVID Search Results."—NPR [1]

Search engines and social media platforms are an indispensable part of our lives; 92% of the adult population relies on them for information with 52% doing this on an average day [320]. Despite their increasing popularity, to date, their search, ranking, and recommendation algorithms remain a black box to the users. The relevance of results produced by these search engines is mostly driven by market factors and not by quality (*fairness*, *credibility*, and *representativeness*) of the content of those results [397]—a fact most people are unaware of [262]. There is no guarantee that the information presented to people on online platforms is credible. The repercussions of users' exposure to fabricated information in search results combined with their unwavering trust in online platforms could be enormous. Previous research has already demonstrated that online misinformation can manipulate individuals' social, political, and health-related behavior, leading to inaction and detachment [124]. When citizens

base their decisions on inaccurate information, it could not only pose a threat to democratic processes [241] but also impact their health and well-being [232, 247, 402]. Consequently, it is crucial to prioritize the investigation of algorithmically curated misinformation and develop effective long-term defenses against it.

This dissertation research aims to address these concerns through two primary objectives. First, it seeks to understand the role played by algorithms employed by online platforms in amplifying online misinformation. Second, it aims to design defenses against online misinformation by incorporating human-centered insights from stakeholder groups like fact-checking organizations and news agencies. My work acknowledges the complex multifaceted nature of online misinformation and recognizes that in order to design effective defenses against online misinformation, it is crucial to leverage the expertise and insights of stakeholder groups who are actively combating misinformation in the real world. To this end, my research delves into three interconnected threads, each focusing on a distinct aspect of the misinformation problem: online algorithms, fact-checking, and the design of systems to combat misinformation. The first thread of my research addresses the algorithm problem by investigating and auditing the online platforms to understand the role of algorithms driving these platforms in surfacing misinformative content to users **(Chapters 3, 4, and 5)**. The second thread focuses on supporting online fact-checking as a means to combat misinformation. Here, I explore how fact-checking is performed in the real world, identifying stakeholder groups involved in the process, and uncovering the technical, policy, and information barriers to fact-checking online misinformation **(Chapter 6)**. The final thread of my research aims at designing and building an online system that helps fact-checkers in monitoring online platforms for algorithmically curated misinformation **(Chapter 7)**. In the rest of the introduction, I will elaborate on the definition of online misinformation and then will briefly outline each of my research arcs.

1.1 Study Context: Misinformation

The research community has referenced online misinformation with different names and definitions. A few popular characterizations include "fake news" [184, 198], "hoaxes" [243], "rumors" [150, 321], "conspiracy theories" [68, 341], "information credibility" [84, 281] and "perceived accuracy" [78, 308]. In my dissertation, especially the first thread of my research, I focus on the conspiratorial aspect of misinformation and

use these terms interchangeably. Conspiracy theories are narratives that embody the belief that secret and influential organizations are behind the occurrence of a particular event [439]. Note that conspiracy theories are not always false. There have been several cases in the past where conspiracy theories turned out to be true (for example, Watergate Scandal [357] and Project MKUltra [421]). To differentiate true conspiracy theories from false ones, I depend on the theory of social constructionism where a fact is only considered "true" if its claim is widely cited, replicated, and accepted without contest [246]. For the purpose of my audit research, I focus only on conspiratorial topics whose mainstream view of reality is known—for e.g., "vaccines do not cause autism". The mainstream perspective of such theories is either backed by expert authorities or scientific research and is widely accepted by a large number of people. At present, "what the majority of people believe in" is our best effort in determining the truthfulness of conspiracy theories. I acknowledge that the truth value may change in the future if new information is available.

For each of the topics under investigation, I operationalize the credibility assessment task involving annotation of social media content (e.g. YouTube videos, Amazon products, etc.) using social epistemology [166], as done by prior research [279]. According to social epistemology, the consensus among individuals can be considered one of the ways for determining truth [166]. In this perspective, if a majority of individuals hold the same belief, it is often seen as an indicator of truth. This approach is based on the idea that collective agreement can be a reliable measure of the truth. Thus, while getting credibility annotations from multiple individuals, I assigned the credibility label using the majority rule. However, for the third thread of my research, where I built a fact-checking system to assist fact-checkers with misinformation discovery, I leave the determination of the veracity labels to the fact-checkers and their organizations' established processes.

1.2 Research Arcs

In this section, I provide an overview of each thread of my dissertation research.

1.2.1 Auditing online platforms to measure the prevalence of algorithmically curated misinformation

Search engines and social media platforms are the primary gateways of information. However, algorithms powering these platforms optimize for relevance and engagement with no regard for the credibility of the information while presenting content to the users. My research, first, aims to understand the role of algorithms driving these online platforms in surfacing misinformation. I have designed audit methodologies to determine the effect of user features and user activities on the amount of misinformation surfaced by online platforms in searches and recommendations. Using this methodology, I conducted an exhaustive set of carefully controlled experiments to audit social media search interfaces such as YouTube and Amazon. Through my audits, I identified the conditions under which algorithms present misinformative content to users as well as vulnerable user populations who could be targets for certain misinformative topics on online platforms. I briefly describe the three audit studies that I performed under this thread below.

1.2.1.1 Auditing YouTube for perennial and demonstrably false conspiracy theories

In the first study under this research thread, I conducted audit experiments to investigate whether personalization (based on age, gender, geolocation, or watch history) contributes to amplifying misinformation on YouTube (**Chapter 3**). After shortlisting five popular topics known to contain misinformative content (Chemtrails, Flat Earth, Vaccine Controversies, etc.) and compiling associated search queries representing them (via Google Trends and YouTube auto-complete suggestions), I conducted two sets of sock-puppet audits—Search- and Watch-misinformative audits. The audits resulted in a dataset of more than 56K videos compiled to link stance (whether promoting misinformation or not) with the personalization attribute audited. The videos corresponded to three major YouTube components: search results, UpNext, and Top 5 recommendations. I found 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 users develop a watch history, these attributes do affect the extent of misinformation recommended to them. For example, I found YouTube recommending misinformative videos to men watching neutral videos about conspiratorial topics.

1.2.1.2 Auditing YouTube for election misinformation

In this study (**Chapter 4**), I conducted a post-hoc audit of election misinformation on the YouTube platform. After the US presidential elections, a lot of conspiracies circulated on YouTube questioning the validity of the election procedures as well as the results of the elections. In response, YouTube established content policies to remove videos promoting election-related falsehoods from its platform and said that such misinformative videos would not prominently surface in its searches and recommendations. In this work, I conducted a large-scale crowd-sourced audit of the YouTube platform to determine how effectively YouTube regulated its algorithms—search and recommendation—for election misinformation. To conduct the investigation, I recruited 99 participants with different demographics and political affiliations who installed *TubeCapture*, a browser extension that I built to collect users' YouTube search results, and recommendations. The extension conducted searches for 88 search queries related to the 2020 US presidential elections and collected up-next recommendation trails—five consecutive up-next recommendation videos—for a set of pre-selected seed videos. I found that YouTube's search results, irrespective of search query bias, contain more videos that oppose rather than support election misinformation. However, watching misinformative election videos still lead users to a small number of misinformative videos in the up-next trails.

1.2.1.3 Auditing e-commerce platforms for health misinformation

In the third study under this research thread, I conducted two sets of algorithmic audits on the Amazon platform to examine the prevalence of vaccine misinformation in its search results and recommendations (**Chapter 5**). First, I systematically audited search results belonging to vaccine-related search queries without logging into the platform—unpersonalized audits. Second, I analyzed the effects of personalization due to account history, where history is built progressively by performing various real-world user actions, such as clicking a product, adding a product to cart, etc—personalized audits. My work provides an elaborate understanding of how Amazon's algorithm is introducing misinformation bias in the product selection stage and ranking of search results across five Amazon filters for ten impactful vaccine-related topics. My analysis of Amazon's product page recommendations suggests that recommendations of products promoting health misinformation contain more health misinformation when compared to recommendations of neutral and debunking products. Through my audit

experiments, I empirically establish how certain real-world actions on health misinformative products on Amazon could drive users into problematic echo chambers of health misinformation.

1.2.2 Identifying ways to support fact-checking online misinformation

In order to design effective long-term solutions against online misinformation, it is important to understand and support the current fact-checking practices. A lot of times, fact-checking is only considered a technical problem concerning methodologies, tools, and algorithms to detect, assess, and verify the accuracy of claims and information. However, fact-checking is a complex socially-situated technical phenomenon involving collaboration among multiple stakeholders. In this thread of work, I highlight both the social and technical aspects of fact-checking (**Chapter 6**). First, I foreground the social aspect—the human infrastructure—of fact-checking by revealing the synergistic collaboration that occurs among various stakeholder groups that work together to accomplish fact-checking work [317]. Second, I highlight the technical aspect—technological infrastructure—which comprises of tools, technology, processes, and policies that support and enable the work of the stakeholder groups. The foregrounding of the infrastructures supporting the fact-checking work helped me in unraveling the technical, policy, and information barriers to fact-checking. Based on my findings, I suggest that improving the quality of fact-checking requires systematic changes in the civic, informational, and technological contexts. For this work, I interviewed 14 fact-checking organizations across 14 continents, enabling me to get a global perspective on the current fact-checking practices and needs of the fact-checking organizations.

1.2.3 Defending against online misinformation via system design

In my third research thread, I design a fact-checking system that caters to the needs of fact-checking organizations. Through collaboration with these organizations, I discovered that monitoring algorithmically driven online platforms still heavily relies on manual efforts by fact-checkers. They spend significant time conducting manual searches on search engines and social media platforms to identify misleading content. Moreover, generating effective search queries to uncover potentially dubious content remains a challenge, often relying on guesswork. This problem is particularly

pronounced on video search platforms like YouTube, where the lack of dedicated monitoring tools further exacerbates the issue. To address these challenges, I partnered with Pesacheck, Africa's largest indigenous fact-checking organization, to develop YouCred. YouCred is an online fact-checking system that automatically generates search queries related to important events and topics of interest to fact-checkers and provides an easy interface to analyze and annotate YouTube videos. To ensure the system met the needs of fact-checkers, I regularly gathered feedback from the Pesacheck team, leading to refinements and enhancements in the system's interface, features, and functionality. The finalized version of YouCred was deployed and evaluated at Pesacheck for nine months. The response from the fact-checking community was positive, with consistent usage observed throughout the evaluation period. The development of YouCred serves as an example of how participatory design methods can bridge the "design-reality gap", aligning the needs of fact-checking stakeholders with the technical systems designed to support their work.

1.3 Contributions and impact

Below I discuss the broader impacts of my dissertation research.

- **Establishing the phenomenon of algorithmically curated misinformation:** My work has resulted in a methodology to audit search engines and social media platforms for misinformation and opened up a new avenue in the domain of algorithmic-audit research. My audit study of the YouTube platform was the first to empirically establish the prevalence of the "misinformation filter bubble effect" revealing how search engines could trap people in echo chambers of misinformation. By conducting an exhaustive list of experiments, my work has demonstrated the distinct characteristics of algorithmically curated misinformation and has quantified its prevalence across multiple attributes such as user features, user actions, search query types, and time variations.

- **Policy implications:** As a sign of direct policy implications of my work, U.S. Representative Adam B. Schiff cited my algorithm audit work on Amazon in his congressional letter where he asked Amazon to address the problem of vaccine misinformation on its platform [347]. Additionally, I was also interviewed by the U.S. House Select Subcommittee staff who wanted to learn about the challenges of COVID-19 and vaccine misinformation in relation to e-commerce platforms.

- **Creation of novel datasets:** The comprehensive audits conducted as part of my thesis resulted in the creation of two valuable and novel datasets. First, the YouTube audit study (**Chapter 3**) produced a novel dataset comprising 56,475 videos, which linked the veracity label of the video (promoting, neutral, or debunking) with the audited personalization attribute. Second, the audit experiments on Amazon (**Chapter 5**) yielded a dataset of 4,997 unique Amazon products distributed across multiple search queries, search filters, recommendation types, and user actions collected over an extensive 22-day audit period. These datasets serve as invaluable resources, facilitating further research and analysis in the field of online misinformation, specifically regarding the impact of algorithmic curation on content credibility.

- **Determining ways to support the online fact-checking process:** Through my research, I actively engaged with 26 individuals from 14 fact-checking teams and organizations representing four continents, to determine ways to support the online fact-checking process. This collaborative effort provided invaluable insights into real-world fact-checking practices, shedding light on the often invisible advocacy, policy, and research work carried out by these organizations. Through this research, I also identified the diverse technical, social, informational, and policy needs of fact-checking organizations across the globe, contributing to a better understanding of their challenges and requirements in combating misinformation.

- **Narrowing the design-reality gap in the development of fact-checking systems:** Through my work, I have tried to bridge the design-reality gap in the development of fact-checking systems. A significant step in this direction is the design and development of YouCred, a fact-checking system that assists fact-checkers in discovering and assessing misinformation on the YouTube platform. This endeavor involved a two-year collaboration with the Pesacheck fact-checking organization, whose members played a pivotal role in shaping the system. To demonstrate the practicality and effectiveness of YouCred, I conducted an extensive nine-month evaluation, showcasing its applicability in combating misinformation and enhancing the fact-checking process.

CHAPTER 2

RELATED WORK

My research focuses on multiple aspects of online misinformation and is informed by a large body of multi-disciplinary research. This chapter provides a comprehensive review of the literature, exploring three primary dimensions through which I investigate the phenomenon of online misinformation: auditing algorithms employed by online platforms, fact-checking online misinformation, and designing for mitigating online misinformation.

2.1 Auditing Online Platforms for Misinformation

2.1.1 Misinformation in algorithmic platforms

Search engines are modern-day gatekeepers and curators of information. Their black-box algorithm can shape user behavior, alter beliefs and even affect voting behavior either by impeding or facilitating the flow of certain kinds of information [117, 134, 238]. Despite their importance and the power they exert, to date, their search results and recommendations have mostly been unregulated. The information quality of a search engine's output is still measured in terms of relevance and it is up to the user to determine the credibility of the information. In recent times, these search engines have been critiqued for promoting misinformative and biased results [411]. For example, researchers found that a significant number of people ended up believing that the Earth is flat after watching recommended videos on Youtube—one of the

most popular video search platforms [297]. Another report outlined that searching for "vitamin K shot" on Google and YouTube returned web pages and videos asking parents to skip the vitamin shot [120]. The top search results also promoted anti-vaccine conspiracies [120]. In another instance, a search for the term "vaccine" on Amazon resulted in pages dominated by anti-vaccination books and movies, some including sponsored posts and ads [145]. All the aforementioned studies provide anecdotal evidence of algorithms playing a role in surfacing misinformation without experimentally quantifying its prevalence. My proposed research aims to fill this gap by applying the audit methodology to empirically establish the conditions under which the algorithms driving online platforms surface misinformation.

2.1.2 Search engine audits

In recent times, search engines have been critiqued for promoting misinformative and biased results [411]. One of the key methodologies used to identify, study, and quantify such bias, discrimination and misinformation is the *audit methodology*. An audit comprises of systematic statistical probing of an online platform to uncover societally problematic behavior underlying its algorithms [344]. Scholars have proposed and used a myriad of audit research methods, including code audits, scraping audits, sock puppet audits, and crowd-sourced audits [344].

A *code audit* requires researchers to get access to the algorithm's code and design in order to analyze it for problematic or harmful behaviour. Such audits are unfeasible for online platforms since their code is proprietary and is not available for public access [305, 306]. Furthermore, such audits could require human experts to understand and untangle the code logic [79]. This method is also only useful in detecting a limited range of problems in algorithmic systems since algorithms do not exist in a vacuum, and they might only show biased behaviour when they act on users' data [344].

A *scraping audit* involves researchers collecting data directly from the web page or via an API. This audit method is not useful in situations where user characteristics (such as gender, age, etc.) impact the algorithmic output [423]. In a *sock puppet audit*, researchers create bot accounts or fake user accounts that impersonate real-life users in order to investigate how an algorithmic system may behave in response to different user characteristics or user actions. This audit method gives researchers the greatest control over experimental variables [423]. However, this method involves injection of false and harmful data into the platform under investigation and necessitates serious ethical considerations. Researchers need to ensure that the actions performed by the

fake accounts do not negatively impact the real users of the platform. Finally, in a *crowdsourced audit*, researchers hire crowdworkers to collect data from the platform in order to test the algorithmic system. Just like sock puppet audits, this method could also inject false and harmful data into the platform. High participant recruitment cost is another limitation of this research design [344].

Using these audit techniques, researchers have investigated several issues pertinent to algorithmically driven online platforms. For example, they have explored the presence of partisan bias in search engine components [202, 275, 330]; investigated representativeness issues, such as racial and gender bias in online freelance marketplaces [189] and resume search engines [93]; the presence of price discrimination and algorithmic manipulation in e-commerce websites [95, 188]; opacity in price surging algorithms used by ride-sharing services [94]; lack of news source diversity in the information returned by search platforms, [393]; and the extent of personalization and localization used by search engines [187, 235]. Yet, auditing online platforms for algorithmic misinformation is practically non-existing. By focusing on auditing online platforms such as YouTube and Amazon for misinformation, my research takes a first step in the direction of auditing algorithms for misinformation.

2.1.3 Methodological Challenges in Audit Investigations

There are numerous methodological challenges while conducting audit investigations. The first roadblock is determining a viable set of search queries that will result in meaningful measurements. Surely, we cannot feed all possible search queries to the system under audit. Researchers have adopted several techniques to compile and shortlist meaningful search queries. For example, to audit Google's Top stories box, researchers selected Trending topics from Google Trends at a fixed time every day and then manually shortlisted the trending queries related to those topics [393]. An audit conducted on Google, Yahoo, and Bing search engines, during the 2016 United States Congressional elections, used the names of electoral candidates as queries [276]. To investigate gender bias in the resume database, researchers used the most commonly searched job titles [93]. To audit for partisan bias in Google search, scholars compiled autocomplete suggestions for multiple root queries related to Donald Trump's presidential inauguration [330]. In my work, I leverage both, queries from Google Trends as well as autocomplete suggestions to ensure that the query set is trending and relevant to the platform under investigation.

The second challenge of audit methodologies relates to carefully controlling the

experimental setup for meaningful audit investigations. These comprise decisions on setting the data collection framework, selecting the components to audit, and controlling for confounding factors or noise. What audit methodologies should one select for conducting the audit? Researchers in the past have used various methods to collect data for the audit experiments [59]. In my work, I have employed both sock-puppet and crowdsourcing methods. For two projects, I manually crafted accounts on online platforms and used automated scripts to collect data so as to have more control over the experiments. For the third project, I recruited individuals who were instructed to install browser extensions to collect data, enabling me to observe algorithmic behavior in response to the complex user histories of real individuals. What components should one select for the audit experiments? Some audit studies focus on one component of the search engine, such as Google's Top stories box [393] or Google's search results [187]). Others focus on multiple components combined, such as various Google search page components including people-ask, news-card, twitter, people-search etc. [332]). My audit studies focused on search results and various recommendations specific to the platform under audit. I also leverage previous literature [187] to control for any confounding factors that could possibly affect the outcome of the experiments.

The third challenge for conducting search engine audits lies in identifying the attributes and actions that could possibly affect the feature one is auditing. Several audit studies have focused on geolocation-based personalization. For example, to investigate the effects of geolocation on web-based personalization, researchers focused on nation-level (randomly selected states in the USA), state-level (counties within Ohio) and county-level (voting districts in Cuyahoga County) locations. They found that personalization in search results increases with physical distance [236]. Audit studies have also investigated the effects of demographics, search-history, click history, and browsing history on Google's web search results as well as prices of commodities on e-commerce platforms [187, 188]. Motivated by these studies, I investigate various user attributes and actions relevant to the platform under audit and determine their impact on the amount of misinformation that gets surfaced on online platforms. The last challenge for conducting online audits relates to properly defining how one is measuring the output label of the phenomenon that is being audited. For example, if a study investigates partisan bias, how do you define and label bias in a valid way? For my work on misinformation audits, I label algorithmic content as promoting, debunking, or neutral based on whether it supports, debunks, or presents general information about the topic of the audit study.

2.2 Fact-checking online-misinformation

With the presence of a vast amount of information online, it is becoming increasingly difficult to judge what to believe or discredit [66]. One of the most prominent approaches to identifying information accuracies on online platforms is fact-checking. Thus, to combat misinformation it is essential to support every aspect—both visible and invisible—of the fact-checking process. In this section, I first present the definition, origin, and evolution of fact-checking (Section 2.2.1). Next, I discuss the literature on the invisible work of fact-checking (Section 2.2.2), and finally the current landscape of research in fact-checking (Section 2.2.3). I show how previous work engages with understanding the fact-checking practices and tools in a limited manner and describe how my research addresses this gap.

2.2.1 Fact-checking: Definition, Origin, and Evolution

The American Press Institute defines the process of fact-checking as "re-reporting and researching the purported facts in published/recorded statements made by politicians and anyone whose words impact others' lives and livelihoods." [133]. One of the early examples of fact-checking emerging as an integral part of journalism was when the *Time* magazine set up a separate research department to objectively verify every printed word before releasing the publication, a phenomenon now known as ante hoc, internal, or in-house fact-checking [119]. The last decade also witnessed the emergence of post-hoc or external fact-checking which consists of publishing an evidence based analysis of claims made in any public text (e.g., news report, political speech, social media posts, etc.) after it is released to the world [174]. Today, fact-checking has emerged both as a principal part of news reporting as well as a separate entity [400]. According to Duke Reporters' Lab, by 2019 there were around 188 active fact-checking initiatives spread across 60 countries [370]. These initiatives have incorporated a range of methodologies and data-driven journalistic practices to not only hold disinformation-spreading individuals and organizations accountable through their fact-checks, but to also disseminate fact-checks in such a way that increases engagement with the public [58, 99]. With the aim of bringing together these fact-checking initiatives and in order to promote common fact-checking standards through a code of principles, the International Fact-checking Network (IFCN) was established in 2015 [315]. Major social media companies such as Facebook and Google have since then partnered with IFCN signatories to debunk false claims surfacing on their platforms [175]. While existing

work describes the evolution of fact-checking from journalism to external fact-checking [172, 177], there are still gaps in understanding how fact-checking is actually practiced, the identity and role of the participating stakeholders and the various collaborations and partnerships occurring in the process. Understanding these aspects is essential to support the process of fact-checking online misinformation. Through my research, I deep dive into answering these missing aspects of the fact-checking phenomenon.

2.2.2 Invisible Work of Fact-checking

Within CSCW, a lot of attention has also been paid on highlighting the invisible or the overlooked work in a process or within an organization [38, 131, 142, 345]. Invisible work can include situations where the person performing the work is visible but some of the work they perform is "functionally invisible or taken for granted" [372]. Such work remains hidden in the background but is essential for the collective functioning of a workplace [289]. It often includes informal work practices such as informal conversations, operational and maintenance work, etc [289, 396, 401]. There are also situations where the person performing the work itself is invisible, such as service, design, or domestic work [266, 368]. In certain complex environments (e.g. hospitals), both visible and invisible work practices can take place simultaneously [372]. In a similar vein, fact-checking is a complex ecosystem that includes somewhat *visible* editorial and investigative work and *invisible* advocacy, policy, and research work. Most of the prior research has looked at fact-checking as a process to debunk misinformative claims[49, 173, 174, 362]. Scholars have mostly engaged with the role of stakeholder groups such as fact-checkers and editors in supporting the fact-checking work [49, 176, 362]. I add to the existing literature by not only expanding on the previously reported roles of fact-checkers and editors but also identifying other stakeholder groups, such as investigators and researchers, and advocators whose roles remain invisible and unexplored by prior research. I shed light on the invisible work that fact-checking organizations are doing to improve the availability and quality of the information in their country. By rendering visibility to the invisible work in fact-checking, I hope to foreground the challenges faced by stakeholders involved in every step of the fact-checking work. This would in turn open avenues for future research supporting various aspects of the fact-checking process.

2.2.3 Current Landscape of Research in Fact-checking

Past research studies on fact-checking have primarily focused on automating multiple stages of the fact-checking process [72, 88, 158, 191, 228, 356], determining the perception and believability of fact checks [44, 149, 295] and construction of fact-check databases [239, 385, 410]. Scholars have adopted several approaches to determine the veracity of content, such as use of knowledge graphs [359], crowd-sourcing [193, 233], deep learning models [225], natural language processing techniques coupled with supervised learning techniques [191] and combination of human knowledge and AI [292]. Work in the field of multimedia forensics has also led to the development of content verification tools especially for image and video verification such as Tineye, InVID, etc. [380]. Despite the plethora of automated systems and tools available for fact-checking, our understanding of their usefulness in practice is limited. Furthermore, there is a dearth of scholarly work that engages with the limitations of the current fact-checking tools and practices ([118, 171] are a few exceptions). My research addresses this gap by interviewing the various stakeholder groups involved in the fact-checking process to understand the the technological infrastructure supporting their work including the tools that are actually used by the stakeholders in practice, the limitations of the current tools, and the challenges faced by multiple stakeholder groups.

2.3 Designing for Mitigating Online Misinformation

In response to the prevalence of online misinformation, scholars have proposed multiple solutions to combat online misinformation. The solutions span several approaches, including designing interventions on platforms [70, 97, 104, 156, 220], media literacy programs [85, 113, 213, 391] as well as designing games to help people build resistance to the online fake news [61, 170, 260, 336, 337]. A lot of these existing approaches are aimed to aid users in addressing the credibility of information that they see online [398]. However, scholars have argued that while it is important to add design features that would help people navigate a large amount of online information, the sole burden of determining the credibility of online content should not solely be shifted to users [2]. Thus, researchers are also designing tools and systems to support fact-checkers by automating and scaling various aspects of the fact-checking process (see [287] for a review). Since finding misinformation is one of the most challenging aspects of fact-checking process, several tools have been developed to monitor online platforms. For example, *WhatsApp monitor tool* monitors multimodal content (text, image, video)

that is being posted and shared on a set of public WhatsApp groups and displays the content shared the most number of times by the users [272]. *Watch 'n' Check* allows fact-checkers to monitor trending topics on Twitter and find tweets containing specific keywords. The tool also displays metadata of the tweet along with details about the communities of users sharing the tweet to assist fact-checkers in making better decisions [88]. Crowdtangle allows monitoring of public Facebook groups and pages and is widely used by fact-checkers to find misinformation on the platform [138]. In addition to the externally available tools, some platforms also have internal tools only accessible to partner fact-checking organizations to discover misinformation. For example, Meta has a misinformation monitoring tool, colloquially known as The Queue that surfaces user-submitted and AI-surfaced potentially misleading content on Facebook and Instagram [138]. While there is a lot of research and available tools for platforms like Facebook and Twitter, there is a dearth of external or platform-supported internal tools to monitor misinformation on video search platforms like YouTube. My work fills this gap by designing a monitoring system that assists fact- checkers by suggesting search queries that could lead fact-checkers to potentially misinformative content on the platform.

Apart from platform monitoring tools, research has also concentrated on assisting fact-checkers in determining fact-check-worthy claims in a document [192, 218]. Additionally, a plethora of work has been done to determine the veracity of a given claim [84, 152, 193, 438]. Scholars have also built end-to-end fact-checking tools that automate all aspects of the fact-checking process. *ClaimBuster* monitors live discourses and social media platforms to catch factual claims and matches them with a repository of fact-checks to determine their veracity. For the previously unchecked claims, the tool queries search engines and databases (Wolfram Alpha) to find more information about the claim [191]. from online platforms, stance detection of documents with respect to given claims, extracting evidence, and Despite the plethora of available systems designed to facilitate fact-checking procedures, only a fraction have demonstrated real-world impact [171, 317]. This can be attributed, in part, to their limited applicability to diverse real-world scenarios and inadequate consideration of the distinct needs, knowledge, and resources available to fact-checking organizations [317]. To overcome these limitations and to ensure the utility of fact-checking systems within real-world organizations, it becomes imperative to incorporate the specific requirements, knowledge, and expertise unique to the fact-checking organizations while designing fact-checking systems. My work is deeply rooted in this approach, inte-

grating the needs and feedback of fact-checking organizations throughout the design, development, and evaluation phases.

CHAPTER 3

AUDITING YOUTUBE FOR PERENNIAL AND DEMONSTRABLY FALSE CONSPIRACY THEORIES

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—a key aspect governing the ability to meaningfully participate in public life [161]—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 [344]. For example, multiple studies have audited search engines for the presence of partisan bias [202, 330] and gender bias [93, 117]. 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 [216]), despite popular prediction that by 2022, 82% of internet traffic will come from videos [101]. YouTube has also faced years of criticism for surfacing misinformative content [82, 120, 415]. Critics have gone as far as calling YouTube a *conspiracy ecosystem* [45]. Despite such vehement criticisms, there has been little effort toward quantifying the extent of misinformation in video search platforms or investigating user attributes that might have an effect. What is the effect of attributes, such as user's demographics and geolocation on the amount of misinformation returned and recommended on YouTube? How does it change with user's watch history, where watch history is progressively built by watching videos rife with inaccuracies or videos presenting extensive debunks? This chapter grapples with these questions

and sheds light on the phenomenon of algorithmically surfaced misinformation on YouTube and how that is affected by penalization attributes (gender, age, geolocation, and watch history). I study the conspiracy facet of misinformation and perform audits on trending and perennial misinformative topics that are widely known to be false (details in Section 1.1). In particular, I examine five misinformative topics namely, *9/11 conspiracy theories*, *chemtrail conspiracy theory*, *flat earth*, *moon landing conspiracy theories*, and *vaccine controversies*.

I 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. Both audits control for extraneous factors that can lead to potential errors in the audit data collection. I create more than 150 Google accounts to audit YouTube. The experiments collect 56,475 YouTube videos, spread across five popular misinformative topics and correspond to three major components of YouTube: videos present in *search results*, *Up-Next*, and *Top 5* recommendations.

I 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. 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, I 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.

3.1 Research Questions and Hypotheses

My 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? I formulate the following sub-questions and hypotheses to investigate the effects of each of these personalization attributes.

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?

Users provide their demographic information, including age and gender while signing-up for a new Google account. They use the same Google account for accessing YouTube. Prior studies investigating associations between user demographics and engagement with misinformation have found that the likelihoods for sharing misinformation vary across user groups [183]. For example, adults aged 65 or older were seven times more likely to share articles from fake news domains compared to younger age group users. Another study indicated that women have a higher likelihood of sharing misinformation [98]. Different demographics having different likelihoods of sharing misinformation might imply that certain groups are exposed to more misinformative content than others. Thus, given the interplay between demographic differences and engagement with misinformation, I hypothesize that YouTube's algorithm could indeed be biased, exposing older people and females to more misinformation while presenting content related to the five misinformative topics.
H1a. Older people (50 years or older) will be presented with more misinformative content than younger age groups.
H1b. Females will be presented with more misinformative content than males.

Prior studies have also shown that search algorithms, specifically Google search, leverage user's geolocation information to present personalized search results [187].

Moreover, Google keeps track of the region-based popularity of search topics and search queries through Google Trends data [196]. Hence, I hypothesize that geolocation will exert an effect, which in turn will depend on how popular the misinformative search topic is in that region.
H1c. Regions where misinformative topics are popular (hot regions) will be presented with more misinformative content compared to regions where such topics are rarely searched (cold regions).

While RQ1 investigates the effect of attributes that are directly connected to a user's account, RQ2 delves into the second order effect of a user's accumulated watch history. Hence, in RQ2, I ask:
RQ2 [*Watch Experiments*]: What is the effect of watch history on the stance of misinformative content returned in various YouTube components?
Technology critics have raised concerns on search engines' tendency to create a filter bubble over time by presenting less diverse and more attitude confirming search results and recommendations [304, 373]. Some media reports have gone so far as to claim that YouTube recommendations drive users down the conspiracy rabbit-hole by recommending increasingly more pro-conspiracy theory videos [335]. Hence, I hypothesize:
H2. Watching more videos belonging to a particular misinformative stance (promoting, neutral or debunking) leads YouTube's search and recommendation algorithm to present more videos reflecting that particular stance to users.

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?

Some misinformative topics are more popular than others. For example, topics like *vaccine controversies* have been widely discussed in the popular media. In the last few years, several social media platforms received backlash for harboring anti-vaccination content [154, 269]. At the beginning of 2019, a handful of them, including YouTube, pledged to take measures against vaccine misinformation [351, 379]. Does that indicate

that YouTube's algorithm will present less misinformative content for such topics, given I performed the audit experiments in the middle of 2019? I hypothesize that when attention received by misinformative topics varies, the amount of misinformative content presented by topics will also vary.
H3. The amount of misinformative content returned will differ across misinformative topics.

3.1.1 Five Misinformative Topics: Demonstrably False and Perennial

In this research, I focus on five topics namely, *9/11 conspiracy theories*, *chemtrail conspiracy theory*, *flat earth*, *moon landing conspiracy theories*, and *vaccine controversies*. All these topics are demonstrably false, perennial, and denied by authoritative sources or backed by scientific research. I now describe each topic and demonstrate how these are demonstrably false and perennial.

3.1.1.1 9/11 misinformative topic

There are several conspiracy theories surrounding the *9/11* attacks [291]. Some of them claim that authorities had foreknowledge of the attacks and that they deliberately aided the attackers. Few attribute the collapse of the Twin Towers to a controlled demolition or explosives [291]. Possible motives for these theories involve justification of the Iraq and Afghanistan invasion by the U.S. Government. Other theories assert that attacks were financed by Saudi Arabia's Royal family or were orchestrated by the Israel Government or Pentagon was hit by a missile launched under the orders of the U.S. Government [237]. All these accounts have been denied by authoritative sources and expert analysts [374]; hence the theory is demonstrably false. Yet, a New York Times poll conducted on 1,042 individuals revealed that 16% US adults do not believe in government's account of 9/11 attacks and 56% believe that the government is hiding something from them [387]. These statistics reveal that the theory is still persistent, despite being false.

3.1.1.2 Chemtrails misinformative topic

Chemtrails conspiracy theories claim that long lasting condensation trails, also known as Contrails, left by air-crafts and rockets in the sky are composed of harmful chemicals. The theories blame United States Air Force (USAF) for spraying these harmful chemicals with the intention of altering the weather, controlling the population and

causing diseases. National Oceanic and Atmospheric Administration (NOAA) has constantly denied such allegations, citing research that has debunked these false claims [294]. Despite the scientific evidence, a recent study done with 1000 subjects found that 10% and 30% of Americans believe chemtrails conspiracy to be "completely" and "somewhat true", respectively [388].

3.1.1.3 Flat earth misinformative topic

The third topic relates to *flat earth* conspiracies. *Flat earth* conspiracy theorists claim that the National Aeronautics and Space Administration (NASA) and government agencies are duping the public into believing that Earth is spherical in shape. Surprisingly, a 2018 survey revealed that only 66% of millennials believed that the Earth is spherical [431].

3.1.1.4 Moon landing misinformative topic

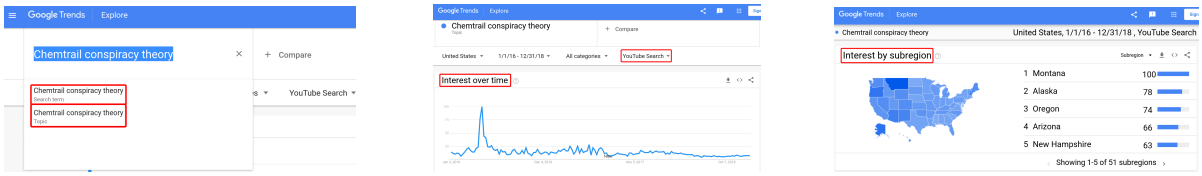
Moon landing conspiracies claim that NASA's Apollo Mission's moon landing was staged by the agency. The theory was denied by NASA [290]. A 500 person poll revealed that 1 in 10 Americans still believe that moon landing never happened [53], justifying the perennial criteria for topic selection.

3.1.1.5 Vaccine misinformative topic

Conspiracy theories related to vaccines are based on the mistaken belief that vaccines contain harmful ingredients that can cause diseases like autism and sudden infant death syndrome (SIDS). Some theories also claim that childhood diseases can be automatically cured by the human body's immune system and thus, vaccination is not required. Such claims are denied by the World Health Organization among other authoritative sources and several scientific research [298, 426]. Yet, a recent survey conducted with 2000 participants revealed that 45% of American adults doubt vaccines [378]. I discuss how I empirically selected these five misinformative topics in detail in Section 6.2.

3.2 Methodology

Here, I first present the methodology for compiling high-impact misinformative queries, the design, and implementation of the audit experiments, the steps for col-



(a) **(b)** **(c)**

Figure 3.1: (a) Google Trends allows users to specify search query as either a topic search or a term search. (b) Interest over time graph. (c) Popularity of *chemtrail conspiracy theory* topic in YouTube searches in the United States between January 1st, 2016 and December 31st, 2018. Color intensity in the heatmap is proportional to the topic's popularity in that region.

lecting audit data, including components of YouTube's Search Engine Results Page (SERP) and video pages, and the qualitative coding scheme for determining the stance of the returned videos.

3.2.1 Compiling High Impact Topics and Queries

My selection methodology to identify relevant and impactful *misinformation search topics* and *queries* comprises three key steps.

3.2.1.1 Selecting misinformative topics via Wikipedia and related research:

I curate a list of relevant misinformative topics (see Table 3.1) by referring to Wikipedia pages on conspiracy theories [6, 419] (e.g., 9/11, chemtrails, sandy hook, pizzagate conspiracy, etc.). I also refer to past studies that examine misinformation and conspiratorial phenomena in online communities [342, 425]. From this list, I exclude topics whose "truth" value is uncertain, that is, topics for which I was either unable to determine the mainstream perspective or the mainstream perspective is not backed by authoritative voices or scientific research. I manually identify and eliminate such topics. For example, I removed "Malaysian Airlines Flight MH370" topic since official investigations about the flight's disappearance have presented inconclusive reports [64, 245, 420]. Next, I 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.

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

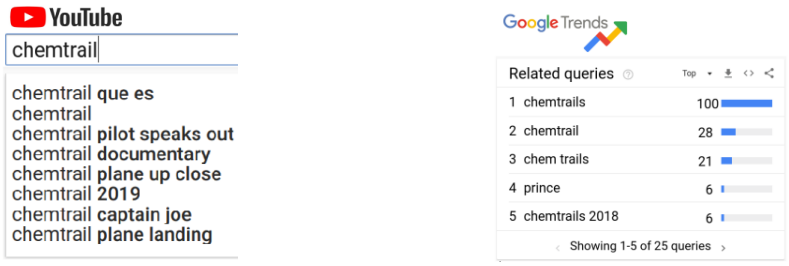
Table 3.1: Seed query, hot & cold regions, and sample search queries for the five misinformation search topics.

3.2.1.2 Selecting high impact misinformation search topics via Google Trends:

Google Trends (Trends for short) is a good indicator for real-world activities impacting a large number of people [128]. Trends also provides interest data across different Google search services including YouTube. Figure 3.1a demonstrates how Trends could be used to search either as a *Term* or as a *Topic*. For example, searching as a *Topic*, *chemtrail conspiracy theory* will give results for several queries related to the topic (*chemtrails, contrails*—a common word used to refer to chemtrails*)*, whereas searching as a *Term* will return results that contain text strings "chemtrail," "conspiracy," and "theory." I opted to search as a *Topic* and selected "YouTube search" as the preferred service (refer to Figure 3.1b). This step discarded a few topics for which no trends data was returned. Next, I compare the *interest over time* plots for all remaining search topics from January 1, 2016 to December 31, 2018 to ensure that the topics have been persistently discussed in the last two years. Then, I select the top 5 topics which represent the most searched topics, resulting in a list of highly impactful misinformative topics. Table 3.1 provides a list.

3.2.1.3 Selecting Search Queries

The next step is to generate a set of queries for each of the misinformation search topics which I can use in the subsequent audit experiments and SERP data collection. I



(a) **(b)**

Figure 3.2: (a) YouTube search's auto-complete suggests 10 trending queries. (b) Google Trends displays the top search queries related to the term or topic entered in the search box.

need to ensure that the query set comprises both relevant and high-impact or popular queries.

I feed seed queries per search topic on both YouTube and Trends. Since the study audits YouTube, query suggestions on YouTube represent the most trending queries searched on the platform, whereas Trends helps identify the most prevalent and impactful queries. YouTube's search box's auto-complete feature suggests 10 popular queries once a seed query is fed into the search box (refer to Figure 3.2a). I add those to expand the query set. Searching on Google Trends as a *Topic* displays top related queries; the number can vary by topic. I also include those in the query set (refer to Figure 3.2b). Thus, the query set comprises queries suggested by both YouTube and Trends platforms.

Next, I manually removed duplicates and replaced semantically similar queries with a single relevant query. I retain the most impactful (trending and most searched) queries by keeping the seed query as well as queries that appear both in the top 5 YouTube suggestions and top 5 related queries list in Trends. I find that shorter queries (length ≤ 4) were better representative of the misinformative topic. Queries comprising more than 4 keywords (for e.g., "the flat earther's $100,000 challenge" and "moon landing press conference analysis") were overly specific. Hence, I only retain more representative generic queries that had a maximum of 4 keywords. The final query set for the *9/11 conspiracy theories* and *vaccine controversies* topics had 11 queries each. Query sets for *chemtrails*, *flat earth*, and *moon landing conspiracy theories* topics had 10, 8, and 9 queries, respectively. In total, I had 49 queries. Table 3.1 presents a sample.

3.2.2 Overview of Audit Experiments

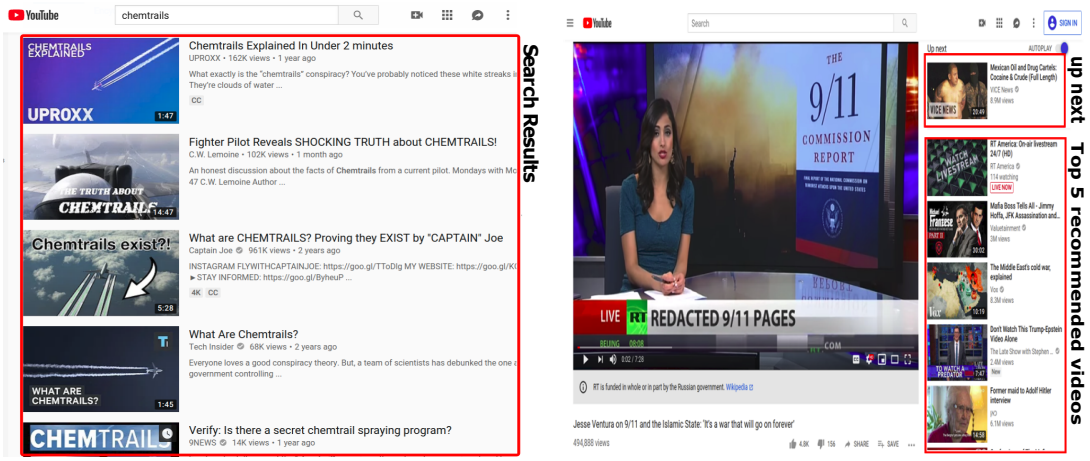
YouTube utilizes age, gender, geolocation, and watch history as features in its recommendation system [106]. To determine if these features amplify the amount of conspiratorial content returned to users, I conduct a series of four audit experiments. The audits collect three primary YouTube components. I annotate the collected videos with stance values: promoting, debunking, or neutral stance towards the topic. Finally, I conduct statistical comparison tests on the annotated data. The audit experiments also control for multiple sources of noise. Unfortunately, in search engine audit studies, differences in search results and recommendations cannot be solely attributed to personalization. Confounding factors (or noise), if not controlled, can also influence the results. For example, users' choice of web browser could impact Google's search results and recommendations and hence could lead to noisy inferences. Thus, following prior search engine audit work [187], I control for browser noise by selecting one single version of Firefox browser for all experiments. Firefox was selected over Google Chrome to avoid the possibility of Chrome browser tracking Google accounts used in my experiments. All interactions with YouTube happened in incognito mode to remove any noise resulting from tracked cookies or browsing history. I also control for temporal effects by performing simultaneous searches. Additionally, all machines used in my experiments had the same architecture, configuration, and version of the operating system (64-bit, Ubuntu 14.04, 3.75GB Ram). This step ensures that there are no temporal effects due to the differences in machines' speeds. In the remaining section, I describe the collected YouTube components and the layout of my experimental setup.

3.2.2.1 YouTube Components

I collect the following components: (a) *search results*. These consist of top 20 videos in YouTube's SERP (Search Engine Results Page) returned in response to a search query. (b) *Up-Next* corresponds to the next recommended video that will be played immediately after the current video finishes, (c) *Top 5* relates to the top five recommended videos on the right of the video page. Figure 3.3 demonstrates the three components.

3.2.2.2 Search Experiments: Auditing with brand new accounts

For the *Search* experiments, I conduct two experiments to test whether demographics (age and gender) and geolocation for a new user (with no prior history on YouTube)



(a) SERP **(b)** Video page

Figure 3.3: Three components collected from YouTube: (a) *search results* from a SERP and (b) *Up-Next* and *Top 5* recommended videos from a video page

have a significant effect on the proportion of misinformative content returned by the platform.

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

I 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 3.2). I create eight different Google accounts—2 (gender values) X 4 (age group values)—each having a unique combination of gender and age. I manually crafted these accounts by following Google's account setup process of adding profile details (age and gender), and including a recovery email and phone verification.

Implementation: Each account is managed by a 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 the effect of geolocation. Figure 4.6 shows the experimental setup. Each bot controlling an account opens Firefox browser in incognito

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	Georgia, Montana, New Jersey, Ohio, South Carolina
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	Georgia, Montana, New Jersey, Ohio, South Carolina
	Watch history	Watch history	Promoting, Neutral, Debunking

Table 3.2: List of user features for the audit experiments.

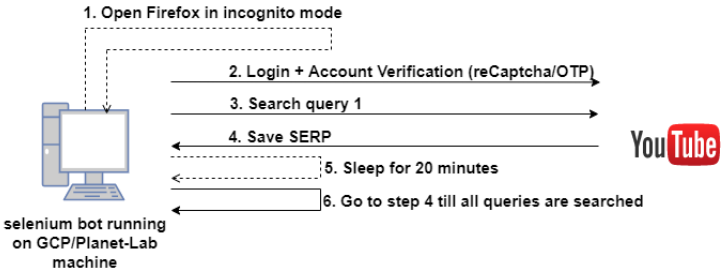


Figure 3.4: Steps performed in *Search* experiments 1 and 2.

mode and logs in to YouTube using that account's credentials. Each bot conducts searches on YouTube's homepage 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 [406]. 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. Prior audit experiments on Google Web Search showed that carry-over effect is observed if the interval between two sequential query execution is less than 11 minutes [187]. I use this value as the benchmark and decide to keep a time interval of 20 minutes between two YouTube searches to control for carry-over effects. I collect SERP data for each of the 49 search queries, scrape these html-based SERPs to extract URLs of the top 20 videos present in the search results.

Experiment 2: Search & Geolocation To study the effect of geolocation, I need to identify physical locations corresponding to each search topic from where automated YouTube searches will be performed. I make use of Google Trend's *interest by sub-region* feature to shortlist locations that have the highest (or lowest) interest corresponding to each topic under audit investigation. I searched Trends 50 times for each of the misinformative search topics with the same parameters (region="US," time="1/1/2016 to 12/31/2018," service="YouTube search"). I calculate the average *interest-by-region* value for each sub-region (i.e. state), shortlist 15 sub-regions with the highest interest scores (referred to as hot regions) and bottom 15 regions with lowest scores (cold regions). Intuitively hot and cold regions are states in the U.S. where the search topic is the most and least popular, respectively. I 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 [312]. For example, for *flat earth* topic, among the 15 hottest sub-regions (e.g. North Dakota, Montana, Oregon, etc.) I selected Montana because of its availability among Planet-Lab active working nodes. Table 3.1 shows the selected hot and cold sub-regions across all topics.

Implementation: For each search topic, I run two selenium bots, each corresponding to either a hot or cold geolocation. The bots run on the virtual machines created on the GCP. These bots

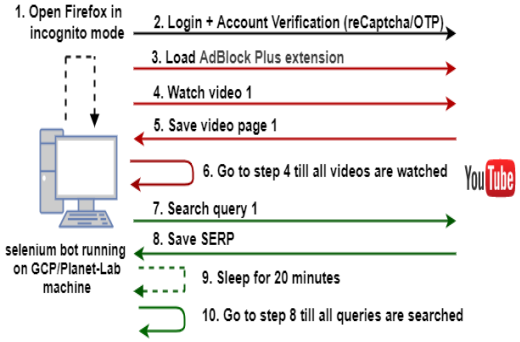


Figure 3.5: Steps performed in *Watch* experiments 3 & 4. These experiments have two phases: (1) watch phase (denoted by →), (2) search phase (denoted by →).

***Watch* experiments, for each misinformative topic:**

Stance	No. of accounts (Demographics)	No. of accounts (Geolocation)	
		Hot	Cold
Debunking (-1)	8	1	1
Neutral (0)	8	1	1
Promoting (1)	8	1	1
Total accounts	24	6	

Table 3.3: Accounts created to execute *Watch* experiments for each misinformative topic. In total, I created 120 (24X5) accounts to run experiment 3 and 30 (6X5) accounts for experiment 4. Here 5 denotes the number of topics.

connect to the Planet-Lab machines deployed in the hot and cold regions (refer to Table 3.1) for that misinformative topic through ssh tunneling. Figure 4.6 presents the steps performed in this experiment. After searching every query, every bot saves the SERP. Later, I scrape all the saved SERPs and extract the URLs of the top 20 videos present in them (i.e. *search results*). After completion of both search experiments (demographics and geolocation), I collected a set of 848 unique videos.

3.2.2.3 Watch Experiments

The goal of the *Watch* experiments is to examine the effect that a user's watch history exerts on the amount of misinformation presented to the user in both YouTube's search and video pages. I also determine how that effect varies with user demographics and geolocation. 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. During the watch phase, after watching every video, I extract the *Up-Next* video and the *Top 5* recommendation components.

Experiment 3: Watch & Demographics. The aim of this experiment is to test the effects in the presence of a user's watch history. Hence, I first need to build the 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. I 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. I obtain the videos from the *Search* experiments. I select the 20 most popular videos for each of the misinformative topics. Popularity is calculated as the engagement accumulated by the video at the time of the experimental runs;

$$\begin{aligned} \textit{Popularity metric (pm)} = \; & view\,count + like\,count + dislike\,count + \\ & favorite\,count + comment\,count \end{aligned}$$

I have released all videos corresponding to each stance (promoting, neutral, debunking) that were used to create watch histories of Google accounts along with their popularity values as the part of the online dataset[1].

Two authors annotated the video collection with stance values: -1 (debunking), 0 (neutral) and 1 (promoting). I describe the qualitative coding scheme and process in Section 4.2.6. Table 3.3 shows the count of accounts created for each misinformative topic for this experiment.

Implementation: The *Watch* experiment for studying the effects of demographics is similar to the *Search* experiment runs. The only difference is 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, either -1, 0, or 1) before conducting any search operation on YouTube. Figure 3.5 presents the steps for the *Watch* experiment.

Experiment 4: Watch & Geolocation. The aim of this experiment is to test the effect of the 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. I create three sets of two Google accounts (see Table 3.3), each corresponding to a hot or cold region (refer to Table 3.1). The three sets build their watch histories following the same steps as in experiment 3.

Implementation: For each search topic, I 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/`

After completion of the experimental runs, I collected 2,479 unique videos from both *Watch* experiments—demographics and geolocation. One author annotated one half of these videos, while the other half was annotated by the second author using the process described in Section 4.2.6.

3.2.3 Annotating the Data Collection

Through the audit experiments, I collected a total of 56,475 videos with 2,943 unique videos. I used an iteratively developed qualitative coding scheme to label the video collection. Qualitative coding is a process of interpreting data and labeling it into meaningful categories. First, two researchers randomly selected 25 videos from the *Search* experiments' data collection, 5 from each topic. Next, six human annotators independently annotated all videos using a basic 3-scale annotation scheme: -1 (debunking), 0 (neutral), and 1 (promoting). All six annotators, including the authors, then discussed their individual annotations and the heuristics followed for the task. After discussions and multiple rounds of iterations, all raters reached a consensus on the annotation heuristics. 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 (details in the next section). For example, the videos could either promote, discuss or debunk the misinformative topic being searched, or they could discuss a different misinformative topic—a topic that the user never searched for. Table 3.4 enlists the annotation values with descriptions and examples. Please note that to curate misinformative topics for the study, I only considered demonstrably false conspiracy theories. But my annotation scheme does not classify videos for veracity, I rather check whether they promote, debunk or discuss a conspiratorial view related/unrelated to the search topic under audit.

3.2.3.1 Annotation heuristics

I annotated videos as "debunking" (-1) when their narrative disputed, derided, or provided scientific evidence against any of the conspiratorial theories related to the particular misinformative topic being audited. For example, the video titled *Bill Maher Throws Out 9/11 Conspiracy Theorists On Live TV* was present in the *Top 5* recommendations while auditing the 9/11 misinformative topic. It mocks people supporting the 9/11 conspiracy theory and hence is annotated as "debunking". Conversely, I

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

annotated videos as "promoting" (1) if they proposed, championed, or substantiated any theory or perspective that promotes inaccurate views related to the topic under audit. For example, the video titled *9/11 truthers attend Treason in America* shows interviews with 9/11 truthers—people who believe 9/11 was an inside job—and hence is annotated as "promoting". I annotated videos as "neutral" (0) when the content of the video presented a general discussion on the topic, without taking stance on conspiracy theories. For example, the video titled *The Howard Stern Show and WCBS-2 On Sept. 11* shows clips depicting damage done to the World Trade Centre after the 9/11 attacks. I marked it as neutral since there is no discussion for and against 9/11 conspiracies.

Annotation values "2", "3", and "4" are similar to values "-1", "0", and "1", respectively, with the difference that they correspond to videos promoting, containing neutral content, or debunking conspiratorial information related to a topic different from the one being audited. For example, consider the scenario where audit experiments of 9/11 misinformative topic returned videos discussing conspiratorial information corresponding to John F. Kennedy's assassination or those pertaining to the Titanic's demise. To illustrate, I list two concrete examples here. Video titled *Did the Titanic Really Sink? The Olympic Switch Theory Debunked* was returned in the *Top 5* recommendations during the *Watch* audits of the 9/11 misinformative topic. The video content refutes the conspiracy theory that claims that the Titanic ship never sank. I annotated it as "debunking misinformation not related to the misinformative topic under audit" (annotation value = 2). In another example, a video titled *JFK coverage 12:30pm-1:40pm 11/22/63* showed news coverage about JFK's assassination without promoting or debunking any false conspiracies. I annotated that video as "neutral video not related to the misinformative topic under audit" (annotation value = 3). On the other hand, a video *Mafia Boss Tells All - Jimmy Hoffa, JFK Assassination and Much More* discusses conspiracy theories surrounding JFK's assassination. I annotated that video as "promoting misinformation not related to the misinformative topic under audit" and assigned an annotation value of 4.

Additionally, I annotated videos as "not related to misinformation" (5) if the content of the video is not related to any misinformative topic. For example, one of the videos in the audit experiment, titled *SHOCKINGLY OFFENSIVE AUDITIONS Have Simon Cowell In A Rage! | ANGRY JUDGES | X Factor Global* is about a reality TV show audition. Since the content does not contain any information related to any misinformative topic, I annotated the video as unrelated to misinformation. Moreover,

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

I annotated non-English videos as "foreign language" (annotation value = 6). I later translated the title, description, and the top few comments of these videos using Google Translate[3]. I then re-annotated them with the appropriate stance value lying between -1 to 5. For example, I re-annotated the Spanish video titled *Las voces del 11S, documental en Español del Canal National Geographic* as "debunking", since the comments within the video indicated that it debunks 9/11 conspiracy theory—the misinformative topic being audited. Finally, videos for which I was unable to assign any annotation value between -1 to 6, I annotated them as "undefined or unknown" (annotation value = 7). For example, the video titled *Ahmed Mohamed's Dad Pushes 9/11 Conspiracy Theories Online* mentions a 9/11 conspiracy tweet. Since the video neither discusses 9/11 events nor takes a stance for or against any conspiracy theory, the coder was unable to decide the annotation value. Because of the confusion it was marked as "unknown". During the annotation phase, I also find that YouTube had taken down 35 unique videos that were captured by the audit experiment. I make an ethical decision to not collect the data or annotate content that was removed by the platform.

After converging on the annotation scale and heuristic, two authors independently coded 158 videos to test for their inter-rater reliability. A high-reliability score (Cohen's Kappa score of 0.80), suggested substantial agreement and offered credence to the annotation heuristic. The authors then split the annotation task of the remaining videos evenly between them. I next develop two scoring metrics to score the amount of misinformation in videos.

3.2.3.2 Normalized scores

The key goal of my audit investigation is to determine whether user activities—search and watch activities corresponding to a particular misinformative topic—leads to more misinformative content, either in the returned search result videos or through the recommended videos. Hence, for downstream analysis, I map the 9-point granular scale (-1 to 7) 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. For instance, if queries for the *9/11* topic result in a video enumerating conspiracies corresponding to missing Malaysian flight 370 (an example from the dataset), then I annotate the video as promoting unrelated misinformation (annotation value = 4) with normalized score = 1. Annotation values of 2, 3, and 4 are mapped to -1, 0, and 1, respectively, while 5

[3] `https://translate.google.com/`

and 6 are treated as neutral (see Table 3.4). I 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, I annotated 2,943 unique videos with 501, 1,980, and 462 videos marked as -1, 0, and 1.

3.2.3.3 SERP-MS Score

I develop a scoring metric **SERP-MS** (SERP Misinformation Score) that captures the amount of misinformation while taking into account the ranking of search results. SERP-MS = $\frac{\sum_{r=1}^{n}(x_i * (n-r+1))}{\frac{n*(n+1)}{2}}$; where r is the rank of the search result and n is the number of search results present in the SERP. I only consider the top 10 search results for computing SERP-MS. Thus, SERP-MS is a continuous value ranging between -1 (all top 10 videos are debunking) to +1 (all top 10 are promoting).

3.3 Results

In this section, I analyze the collected and annotated audit data to investigate my research questions and hypothesis (refer to Section 3.1). The goal is to determine the effects of personalization attributes on the amount of misinformation returned in both *Search* and *Watch* experiments. Recall that, among the three YouTube components (*search results*, *Up-Next*, and *Top 5* recommendations), I can only collect *search results* for *Search* experiments. On the other hand, I collect all three components for *Watch* experiments. A test of normality reveals that the data is not normally distributed and the samples have unequal sizes. Hence, I opt for non-parametric tests. For all pairwise comparisons, I use Mann-Whitney U test. To perform multiple comparisons, I use Kruskal Wallis ANOVA followed by post-hoc Tukey HSD[4]. I report results using both normalized and SERP-MS scores. Note that the SERP-MS score is only calculated for the *search results* component.

3.3.1 RQ1: Effect of demographics and geolocation

In the first research question, I investigate the effect of demographics (age and gender) 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

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

Feature	Topic	Stance	Comp.	Statistical Tests	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)
	Vaccine controversies	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 conspiracy theories	N	Up-Next	MW U=2720, p=0.03	F > M
	Vaccine controversies	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 conspiracy theories	P	Top5	MW U=4137.5, p=0.02	Hot > Cold

Table 3.5: 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.

progressively by watching either promoting, neutral or debunking misinformative videos.

RQ1a [*Search* experiments]: How are *search results* affected for brand new accounts? I 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) when comparing using normalized scores. Use of SERP-MS score also shows non-significant results. Thus, *H1a*, *H1b* and *H1c* are not supported demonstrating that age, gender and geolocation do not have an impact on the amount of misinformation returned in search results for users who have newly created their YouTube accounts.

RQ1b [*Watch* experiments]: How are *search results*, *Up-Next*, and *Top 5* recommendations affected, given accounts have a watch history? I find that age has a significant effect for only two comparisons (refer Table 3.5). In both cases, older people do not receive more misinformation than the other younger age groups. Thus, *H1a* is rejected. Next, I find that gender has a significant effect across five comparisons involving certain combinations of search topics, watch stance, and YouTube components. Out of the five comparisons, *H1b* is supported for one case, where female accounts watching neutral moon landing videos receive more misinformation in their *Up-Next* component than corresponding male accounts watching the same videos. In all other significant

Component	Topic	Test	Mean Diff (post-hoc)
Search Results	Vaccines controversies	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 conspiracy theories	KW H(2,2911)=186.68, p=2.9e-41	P >N & P >D
	Chemtrail conspiracy theory	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 conspiracy theories	KW H(2,3005)=17.18, p=0.0002	P >N & D >N
	Vaccines controversies	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 conspiracy theories	KW H(2,487)=60.12, p=8.8e-14	P >N & P >D
	Chemtrail conspiracy theory	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 conspiracy theories	KW (2,606)=5.99, p=0.049	D >N
	Vaccines controversies	KW H(2,600)=66.86, p=3.0e-15	D >N >P

Table 3.6: 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.

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. Table 3.5 presents all the significant results.

I find that *H1c* holds only for the *Top 5* recommendations of *moon landing* topic. Accounts that watch promoting *moon landing* videos from Ohio (hot geolocation, region with the most interest) receive more promoting videos in their *Top 5* than those who watch the same videos from Georgia (cold geolocation or region exhibiting lowest interest in the topic). For other topics, geolocation did not have any significant effect on the amount of misinformation presented in *search results*, *Up-Next* and *Top 5* recommendations.

3.3.2 RQ2: Effect of watch history

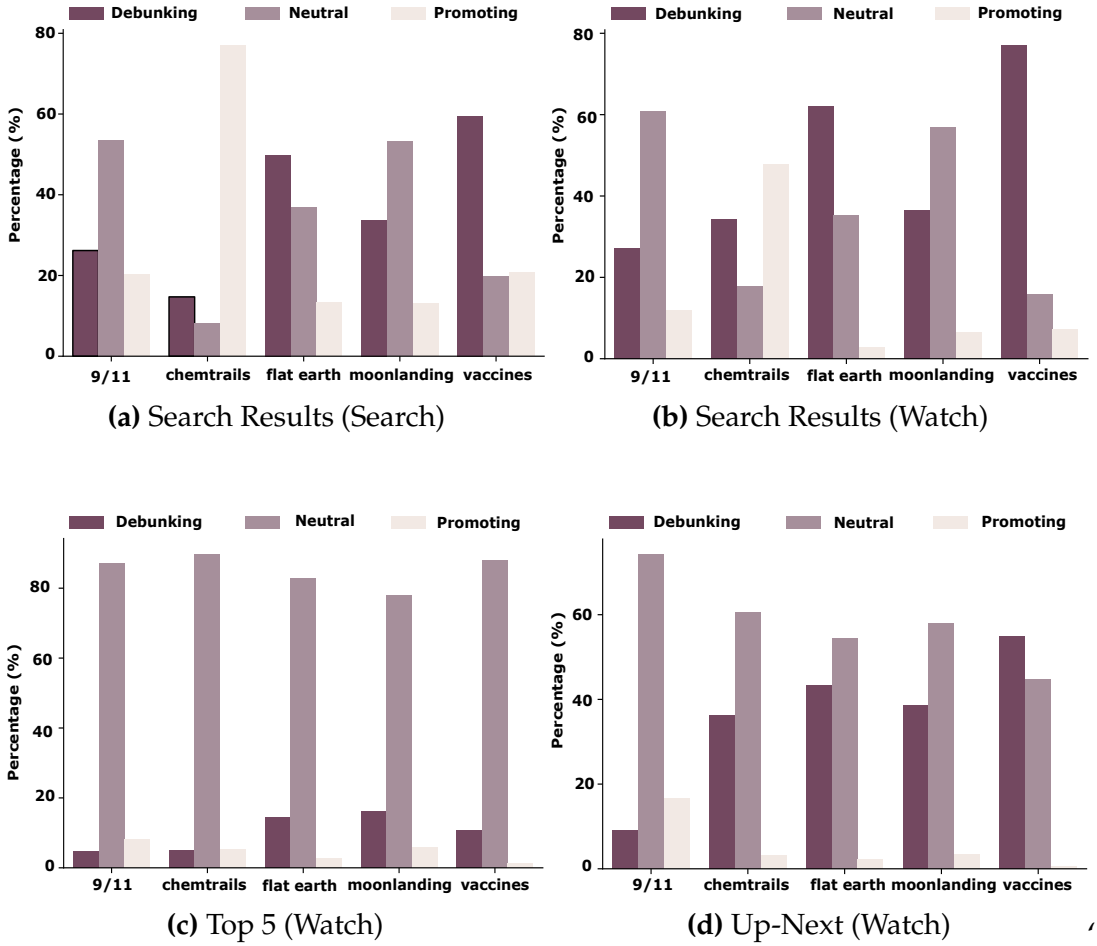
Next, I explore the effect of watch history on the amount of misinformative content returned in the three YouTube components of interest. Note, that RQ2 only applies to the watch experiment, where an account has already built its watch history. Table 3.6 presents only the significant results. I discuss a handful. Statistical tests performed using SERP-MS did not give any significant results. Note that I apply this metric only on the *search results* component. Using the normalized score metric, I find that *H2* only holds for *search results* corresponding to the *vaccine controversies* topic (Kruskal

Wallis H(2,6517)=6.2953, p=0.0429). This indicates that a user's previous watch history only affects the misinformative stance of videos presented in *search results* of the aforementioned topic. 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.

Next, I find that watch history has significant effects on the 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. Thus, *H2* is supported for *Up-Next* and *Top 5* recommendations for all topics. I discuss the post-hoc test results for *vaccine controversies* and *chemtrail conspiracy theories* topics. Post-hoc tests for the *vaccine controversies* topic reveal that 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 presented with more promoting content (Kruskal Wallis H(2,6517)=6.29, p=0.04). Post-hoc tests for *chemtrail conspiracy theories* topic demonstrate that accounts that watch videos promoting chemtrails conspiracies receive more promoting videos in their *Top 5* (Kruskal Wallis H(2,2845)=73.20, p=1.3e-16) and *Up-Next* (Kruskal Wallis H(2,5709)=16.12, p=0.0003) video recommendations than those who watch neutral and debunking videos respectively. Whereas accounts that watch neutral chemtrails conspiracies receive more promoting videos in their *Top 5* compared to those who watch debunking videos of chemtrails. Table 3.6 lists the results for the remaining topic comparisons.

3.3.3 RQ3: Across topic differences

While in RQ1 and RQ2 I studied the effects of personalization attributes on the amount of misinformation presented to users in various YouTube components, in RQ3 I investigate whether misinformative content presented to users differ across the five misinformative topics.



(a) Search Results (Search)

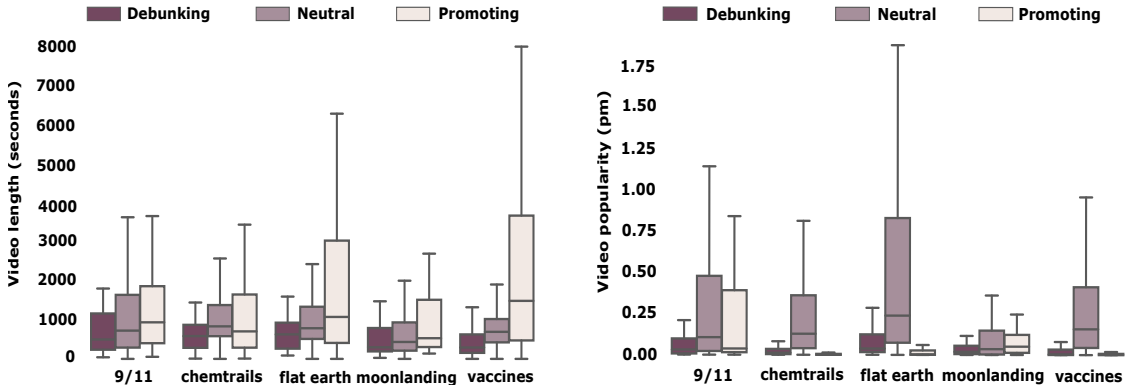
(b) Search Results (Watch)

(c) Top 5 (Watch)

(d) Up-Next (Watch)

Figure 3.6: RQ3: Percentages of video stances for each topic.

RQ3a [*Search* experiments]: How does misinformative content present in *search results* of brand new accounts differ across topics? Figure 5.12c shows the proportion of promoting, neutral, and debunking videos across all topics in *Search* experiments. I find that *H3* is supported for *search results* of brand new accounts. Comparing both normalized scores (Kruskal Wallis H(4,1943)=467.29, p < 7.9e-100) and SERP-MS (Kruskal Wallis H(4,98)=51.1, p < 2.1e-10) across topics show that the amount of misinformation significantly differs among topics. Post-hoc comparisons using Tukey HSD (on both score metrics) reveal that the *chemtrail conspiracy theory* topic harbors significantly more misinformative *search results* compared to all other topics. Figure 5.12c also demonstrates the largest amount of promoting videos in the *chemtrails* topic. I discuss the possible reasons for this occurrence in Section 3.4.2



(a) Video length **(b)** Video popularity

Figure 3.7: Box plots of (a) video length in seconds and (b) video popularity (*pm*) for each stance under each 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 3.6b, Figure 3.6d and Figure 3.6c 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. *H3* is supported for all the three YouTube components for accounts having a watch history. Comparing both normalized scores and SERP-MS across topics, show 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). Recall that SERP-MS is applicable only for the *search results* component. Post-hoc comparisons using Tukey HSD reveal that *chemtrail conspiracy theories* has significantly more misinformation in its *search results* compared to all other topics. Figure 3.6b exhibits the largest amount of promoting videos on that topic. 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 3.6c and 3.6d.

3.3.4 Analyzing Video Length and Popularity

Analyzing video length, I observe that promoting videos are longer than neutral and debunking videos across all misinformative topics, except *chemtrails conspiracies* where they are slightly shorter than neutral videos and longer than debunking ones

(see Figure 3.7a). For all topics, debunking videos are the shortest compared to other stances. I also observe that neutral videos are the most popular (see Figure 3.7b), where popularity is calculated using the *popularity metric (pm)*. Topic *9/11* has more popular videos compared to other topics. On the other hand, for topic *moon landing*, popularity of videos under each stance is almost the same. Although the percentage of videos promoting *chemtrails conspiracies* is highest when compared to other misinformative topics, they are the least popular videos.

3.4 Discussion

3.4.1 Effect of demographics and geolocation on misinformation

Modern search engines filter, rank, and personalize results before presenting them to a user. These information retrieval systems make decisions about the relevance of results without considering accuracy and credibility—a fact most people are unaware of [262]. Motivated by several media reports pointing out the prevalence of misinformation in search spaces, I audited YouTube to empirically determine the extent of conspiratorial content present in its search and recommended video results. I investigate the role played by personalization attributes (age, gender, geolocation, and watch history). The analyses show little evidence to support that user demographics and geolocation play any role in amplifying misinformation in search results for users who have newly started their search journey—those with brand-new accounts. On the contrary, once they have a watch history, I find that demographics and geolocation attributes do exert an effect. However, this effect pertains to only certain combinations of personalization attributes and varies with the topic under audit investigation. I saw significant gender differences in 8 comparisons and in all but one case men (account gender set to "male") were recommended more misinformative videos. Perhaps more surprisingly, in 4 of these cases, men were watching neutral videos and yet ended up with significantly higher misinformative video recommendations. While I do not know why YouTube's algorithm showed this behavior, the observed gender-based differences have important societal implications, especially for certain misinformative topics, such as *vaccine controversies*. For example, a survey of 2,300 people in the United States revealed that the percentage of male anti-vaxxers is more than females [274]. Therefore, recommending videos that promote misinformative topics to men can inflict more harm by reconfirming their pro-conspiracy beliefs. Moreover, recommending

promoting videos to men who are drawn to neutral information and have yet not developed a strong pro-conspiracy belief towards the topic is even more problematic because it might increase their chances of forming pro-conspiracy beliefs.

3.4.2 Effect of watch history on misinformation

One of the goals of the audit investigations was to verify several anecdotal claims criticizing YouTube for surfacing misinformative content in its recommendations [82, 254, 415]. These claims accused the platform of driving users into a misinformation rabbit hole—a phenomenon where people watching videos promoting misinformation are presented with more such videos in the search results and recommendations. Contrary to these blanket claims, I observe variability in YouTube's behavior in presenting recommendations to accounts having a watch history across different misinformative topics. Comparing the stances of the annotated data obtained from the search results of accounts with a watch history shows that YouTube's search algorithm fares well for the flat earth and vaccine topic. On the other hand, I witness a large proportion of videos promoting misinformation for the *chemtrails* topic (refer to Figure 3.6b). This observation can be attributed to YouTube's recent effort to censor misinformative content belonging to select search topics. In an announcement to the public, the platform pledged to reduce misinformative content belonging to topics like 9/11, flat earth and medical misinformation [379]. Thus, I believe the percentage of *search results* promoting these misinformative topics is less compared to other topics like *chemtrail conspiracies*.

The audits reveal that people who watch promoting videos for certain misinformative topics (for example, 9/11 conspiracies) are recommended more of such videos in their *Up-Next* and *Top 5* recommendations compared to those who watch neutral or debunking videos. These findings indicate that the recommendation algorithm is biased towards the stance of videos watched by the user for certain misinformative topics (refer Table 3.6). In another observation I find that for users watching videos on the vaccine topic, both *Top 5* and *Up-Next* recommendations return a negligible proportion of videos promoting "vaccine hesitancy", 1.2% and 0.5% respectively. Statistical tests reveal that people watching promoting anti-vaccination videos receive more debunking videos in their recommendations compared to people who watch neutral or debunking videos. However, a filter bubble effect still exists for the *search results* component, where people watching anti-vaccination videos are presented with more such results. This variability in YouTube's behavior across search topics suggests that

YouTube is modifying its search ranking and recommendation algorithms selectively, handpicking topics that are highlighted by media reports and technology critics (e.g. reports around anti-vaccine video recommendations). These observations are concerning, since all misinformative topics are high impact, popular and perennial and hence are likely to affect a large population of users' search experiences. My findings serve as an important call to action for YouTube to develop a more universal approach that offers a comprehensive solution to the problem of misinformation.

3.4.3 Tackling search engine enabled misinformation

Complete eradication of misinformation from YouTube 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. Which misinformative topics are a threat to public well-being? While "vaccine hesitancy" is now one of the top 10 global threats of 2019 [299] and has led four European nations to lose their "measles free" status [62], seemingly harmless pizzagate conspiracy led a man to fire shots in a pizza parlor [386]. I recommend that YouTube should identify high-impact and popular misinformative topics. My 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.

But is censoring the misinformative content enough? The audit experiments reveal that YouTube recommendations are still biased towards the misinformative stance of videos watched by a user. Given that almost 500 hours of content is uploaded to YouTube every hour [186], censorship might not be a comprehensive solution to fix this algorithmic bias. There is a need to break the filter bubble effect by recommending debunking videos to people who watch videos promoting misinformative content. YouTube can start by identifying and modifying recommendations of vulnerable populations who could be targets for certain misinformative topics. The audit experiments revealed one such demographic. For example, I found YouTube recommending promoting videos to men who watched neutral misinformative videos.

The audits also revealed variability in YouTube's behavior toward certain misinformative topics—an indication of a reactive strategy for dealing with misinformation. I recommend the platform 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". Given the complexity of algorithms used by search engines and the interplay between the data and algorithm, even an expert in the area might not be able to predict algorithmic output [343]. Thus, there is also an inherent need for platforms to conduct audit studies that can help reveal biases present in their algorithm.

While I discussed some nascent steps that YouTube can take towards eradication of misinformation from its platform, this feat cannot be achieved without having proper content policies and infrastructure in place. Currently YouTube's community guidelines do not disallow misinformative content [434]. There is a need to have appropriate policies in place that not only prohibit posting misinformative content on the platform but also ensure that posting advertisements on misinformative videos is not financially incentivized. The challenge of having appropriate infrastructure to implement these policies still remains.

3.5 Limitation and Future Work

This study is not without limitations. I do not perform repeated searches of the search queries over several days which is essential to study the longitudinal effect of personalization. I plan to conduct continuous audit runs with repeated searches in the future. I also tested the effect of the geolocation feature only for regions within the Unites States, but conducting audits over a global scale is a fertile area for future endeavors. The *Search* and *Watch* audit runs had a gap of three months. Thus, I do not perform any comparisons between the *search result* components of the two audits. I do not take into account the stance of a search query and how that affects the search results. I make this conscious choice because my methodology for compiling high-impact search queries, by definition, focuses on realistic searches that were *most used* by real users on YouTube.

Identifying videos that promote conspiracies and inaccurate content or those that debunk them is a challenging task. To make such distinctions with high precision, I used qualitative coding to annotate videos. In addition to the video content, I referred to metadata attributes, such as video title, description, and user reactions present in the comments section. I found that videos relating to misinformative topics exhibit special characteristics. For example, pro-conspiracy videos are mostly longer while neutral videos are more popular. I believe that such distinctive features along with features used in the manual annotation process can be leveraged to build machine learning models that can identify the stance of videos.

While I audit three major components of YouTube, other components such as home pages and trending sections can also be examined. Auditing search queries presented by YouTube's autocomplete feature for their stance is also left for future investigation. Moreover, understanding how misinformative search results and recommendations affect users' search intent [427, 429] is another compelling avenue for future research.

3.6 Conclusion

In this study, I conducted two sets of sock puppet audit experiments on the YouTube platform to empirically determine the effect of personalization attributes (age, gender, geolocation, and watch history) on the amount of misinformation prevalent in YouTube searches and recommendations. I created bots to impersonate users with specific personalization attributes and built YouTube account history by making bots watch videos of certain stances (promoting, neutral, and debunking). I found that the personalization attributes affect the amount of misinformation in recommendations once the bots develop a watch history. The study also empirically establishes the "misinformation filter bubble effect"—the extent to which personalized search engines could trap people in echo chambers of inaccurate information. I also found that the misinformation filter bubbles do not exist equally for all topics. For example, the study suggests that YouTube is modifying its search and recommendation algorithm for *vaccine controversies* topics where the platform recommends scientific videos to users watching promoting videos. As the research delved into these findings, it also propelled further inquiries. Once YouTube modifies its policies and algorithms for a specific topic, what are the long-term effects of such modifications? How do the algorithms behave with real users with complex user histories? I explore these questions in the next chapter.

CHAPTER 4

AUDITING YOUTUBE FOR ELECTION MISINFORMATION

4.1 Introduction

> "Oregon GOP frontrunner for governor embraces claims of election fraud... said he doubted Oregon's vote-by-mail system"—The Texas Tribune, Feb 11, 2022 [360]
>
> "Election Deniers Go Door-to-Door to Confront Voters After Losses (in US primaries)"—Bloomberg, Aug 23 2022 [63]
>
> "With 10 weeks until midterms, election deniers are hampering some election preparations Some election deniers have "weaponized" against us, one election official says."—ABC News, Aug 30, 2022 [364]

Skepticism around the legitimacy of the US electoral process, which primarily gained momentum during the 2020 US presidential election, had serious ramifications. For example, endorsement of election conspiracy theories was found to be positively associated with lower turnout in the 2021 US Senate election in Georgia [178]. In 2022, the false narratives around the 2020 elections still persist [257, 261] and continue to threaten democratic participation in the upcoming US midterm elections [257, 261]. In the last two years, 19 US states altered voting procedures and enacted laws to make voting more restrictive, creating information gaps and fresh opportunities for election misinformation to emerge and proliferate in the real and online world [261]. Thus, battling election misinformation has never been more important.

Studies show that social media platforms have become important mediums for political discourse [46, 408]. In particular, YouTube—the most popular platform among

US adults [310]—has emerged as a political battleground as demonstrated by the fact that both political parties extensively used the platform for election campaigning [384]. However, the platform came under fire from technology critics for being a hub of electoral conspiracy theories [224, 409]. Given the concern that search engines can play a significant role in shifting voting decisions [134, 135] and can confine users into a filter bubble of misinformation [205], there has been a push for online platforms to enact policies that minimize election misinformation [353]. In response to this push, YouTube introduced content policies to remove videos spreading election-related falsehoods and claimed that misinformative videos would not prominently surface or get recommended on the platform [256, 371, 433, 436]. However, the formulation of policies does not equate to effective enactment [318]. It's evident from the results of two misinformation audits conducted on the platform for the same conspiratorial topics (such as vaccine controversies, 9/11 conspiracies), first in 2019 [205] and second in 2021 [390], both of which found echo chambers of misinformation on the platform. Despite changes to YouTube's misinformation policies in 2020 [382], the authors of the second audit study did not find improvements when compared to the results of the first audit, rather they found recommendations worsening for topics like vaccination. These findings iterate the need to continuously audit platforms to investigate how a platform's algorithms fare with respect to problematic content and how effectively a platform's content policies are implemented [361]. While my previous study audited YouTube for misinformation (Chapter 3), it was conducted using sock-puppets (bot accounts emulating real users) in conservative settings[1] which often do not reflect true user behavior. There is a dearth of crowd-sourced misinformation audits that test the algorithms' behavior with real-world users ([71] is one of the few exceptions). In this study, I fill this gap by conducting a large-scale crowd-sourced audit on YouTube to determine how effectively YouTube has regulated its algorithms—search and recommendation—for election misinformation.

To conduct the audit, I recruited 99 participants who filled out a survey and installed *TubeCapture*, a browser extension built to collect users' YouTube search results, and recommendations. The extension conducted searches for 88 search queries related to the 2020 US presidential elections. I also seeded *TubeCapture* with 45 seed videos with three differing stances on election misinformation—supporting, neutral, and opposing. The extension collected up-next recommendation trails—five consecutive

[1]For example, sock-puppet building account history by watching videos that only promote misinformation.

up-next recommendation videos—for each seed video. *TubeCapture* simultaneously collected YouTube components from both personalized standard and unpersonalized incognito windows allowing me to measure the extent of personalization. This leads me to the first research question:

RQ1 Extent of personalization: What is the extent of personalization in various YouTube components?

> **RQ1a:** How much are search results personalized for search queries about the 2020 US presidential elections and the surrounding voter fraud claims?
>
> **RQ1b:** How much are YouTube's up-next recommendation trails personalized for seed videos with different stances on election misinformation—supporting, neutral and opposing?

I find that while search results have very little personalization, up-next trails are highly personalized. I next venture into determining the amount of election misinformation real users could be exposed to under different conditions, such as following up-next trails for videos supporting or opposing election misinformation.

RQ2: Amount of election misinformation: What is the impact of watching a sequence of YouTube up-next recommendation videos starting with seed videos with different stances on election misinformation (supporting, neutral, and opposing) on various YouTube components?

> **RQ2a:** How much do search results get contaminated with election misinformation?
>
> **RQ2b:** What is the amount of misinformation returned in users' up-next recommendation trails?
>
> **RQ2c:** What is the amount of misinformation that appears in users' homepage video recommendations?

I find that YouTube presents debunking videos in search results for most of the queries. I also observe an echo chamber effect in recommendations where trails with supporting seeds contain more misinformation than trails with neutral and opposing seeds. Since election misinformation is closely entangled with political beliefs with several right-leaning news sources amplifying the claims of voter fraud [77, 264], I also study the diversity and composition of the content presented by YouTube in its various components. I ask,

RQ3: Impact on composition and diversity: What is the impact on content diversity when watching a sequence of YouTube up-next recommendation videos starting with seed videos with different stances on election misinformation (supporting, neutral and opposing)?

RQ3a: How diverse are the search results ?

RQ3b: How diverse are the up-next recommendation trails?

I find that YouTube ensures source diversity in its search results. I also find a large number of impressions for left-leaning late-night shows (e.g. Last Week Tonight with John Oliver) and right-leaning Fox news in users' up-next trails. Overall, my work makes the following contributions:

- I conduct a post hoc audit on YouTube to determine how its algorithms fare with respect to election misinformation; post hoc auditing comprises investigating a platform for a past topic or event which could have a significant impact on citizenry in the present and future. In turn, I am able to test the effectiveness of YouTube's content policies enforced to curb election misinformation.
- I extend prior work on misinformation audits by conducting an ethical crowd-sourced audit to see the impact of performing certain actions on the searches and recommendations of real-world people with complex platform histories instead of conservative settings of sock puppet audits.
- My audit reveals that YouTube search results contain more videos that oppose election misinformation as compared to videos supporting election misinformation, especially for search queries about election fraud in presidential elections. However, a filter bubble effect still persists in the up-next recommendation trails, where a small number of misinformative videos are presented to users watching videos supporting election misinformation.

4.2 Methodology

4.2.1 Developing search queries to measure election fraud based misinformation

The first methodological step in any algorithmic audit is to determine a viable set of relevant search queries that would be used to probe the algorithmic system. For my

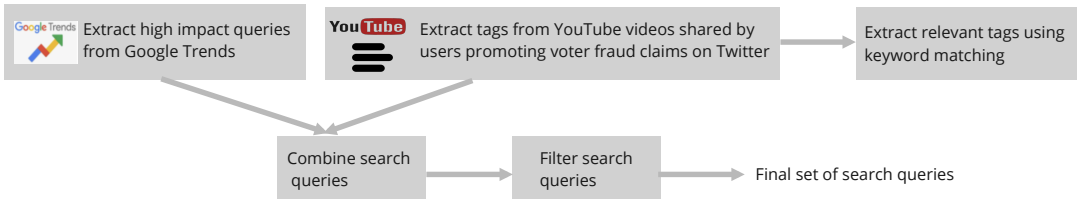


Figure 4.1: Figure illustrating the method to curate search queries for audit experiment

study, I identified search queries that satisfy two properties. First, I select high-impact search queries that were used by people to search about Presidential Election as well as the voter fraud claims about the 2020 elections. Second, I curate search queries that have a high probability of returning misinformative results which would result in meaningful measurements of algorithmically curated misinformation about the audit topic. To compile such queries, I used Google Trends and YouTube video tags (refer Figure 4.1).

4.2.1.1 Curating high-impact queries via Google Trends

First, I leveraged Google Trends which contain Google's daily and real-time search trends data. As the most popular search service, its trends are a good indicator for understanding the real-world search behavior of a large number of people. Using *Election Fraud 2020* and *Presidential Election* as search topics, United States as location, April 2020 to Present as date range, and search service as YouTube search, I extracted the top 15 most and least popular search queries that people used on YouTube. I choose April 7 as the start date since this was the day when Donald Trump made one of his first fraudulent claims about the security of mail-in ballots [208]. I included the most popular queries since they represent the ones that people mostly use to get information on elections. To explore the *data-voids* [167] associated with the audit topic, I also included the least popular search queries to determine if those terms have been hijacked by conspiracists to surface misinformation.

4.2.1.2 Curating misinfo-queries queries using YouTube video tags

Second, I used YouTube video tags that content creators associated with misinformative videos while uploading them on the YouTube platform (see Figure 4.2 for an example). These tags could be thought of as search words representing how content creators would like their videos to be discovered. To extract video tags associated with election misinformation videos, I leveraged a large-scale Voter Fraud 2020 dataset

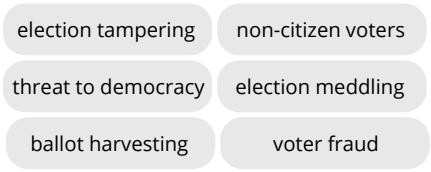


Figure 4.2: List of video tags associated with YouTube video titled `Is Voter Fraud Real?` (video id: `RkLuXvIxFew`) that promotes voter fraud misinformation. Video tags are added by content creators while uploading YouTube videos on the platform. The tags can be extracted from videos via YouTube APIs or third-party tools. I use tags associated with videos shared by users promoting voter fraud claims on Twitter as search queries in the audit experiments.

presidential election 2020
us elections 2020 latest news
election fraud 2020
rigged election
dominion voting exposed
mail in ballots 2020
stop the steal
joe biden voter fraud
usps whistleblower
voter fraud evidence
trump biden general election
dominion voter fraud

Table 4.1: Sample search queries for the YouTube audit

released by Abilov et al [37]. The dataset contains over 12,002 YouTube video URLs that were shared on Twitter by accounts that tend to refute and promote voter fraud claims. I extracted YouTube video tags associated with videos shared by accounts promoting voter fraud claims to probe YouTube (n=200K). To curate a viable number of search queries from the extracted video tags, I employed several steps. First, I manually curated a list of 10 keywords related to elections and fraudulent claims surrounding the elections[2] from the list of keywords provided by Abilov et al [37] as well election 2020 misinformation report produced by the Election Integrity Partnership [147]. Then for each of the keywords, I extracted 15 top and 15 least occurring video tags containing that term. For example, one of the most occurring tags containing keyword *whistleblower* was 'usps whistleblower' while the least occurring tag was 'whistleblower jesse morgan'.

4.2.1.3 Filtering search queries to obtain the final set

I combined search queries obtained from both Google Trends and YouTube video tags in the final query set and employed several filtering steps to obtain a reasonable number of relevant search queries. First, I only kept queries related to election 2020, for example, I kept 'election fraud 2020' and removed 'election fraud 2016'. I removed

[2] *steal, fraud, ballot, elect, seal, dominion, sharpiegate, whistleblower, harvest, and sunrise zoom*

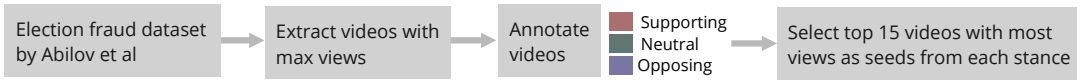


Figure 4.3: Figure illustrating the method to curate seed videos for the audit experiment

Annotation label	Video title	Video id
Supporting election fraud misinformation	Poll worker gives his account of what happened when he tried to monitor the vote in Nevada	4X2V5hPPp6w
	Joe Biden says he's built most extensive "voter fraud" org in history	WGRnhBmHYN0
Neutral	Ex-Trump official shares his prediction if Trump loses 2020	KuqhhrmhfCI
	'Don't be ridiculous': Rudy Giuliani learns about Biden win from reporters	Z0hEFa52Bdo
Opposing election fraud misinformation	Voting by Mail: Last Week Tonight with John Oliver (HBO)	l-nEHkgm_Gk
	Trump and the GOP Still Refuse to Accept Biden's Win: A Closer Look	QoPA3unjQgA

Table 4.2: Sample seed videos curated for the audit experiment.

duplicate and redundant search queries and replaced them with a single randomly selected query. For example, I replaced queries 'voter fraud 2020', 'voter fraud', and 'vote fraud' with 'voter fraud 2020'. I removed queries with lengths greater than five since they were overly specific (e.g. 'we've got pictures of the check stubs paid to people to ballot harvest'). I also removed queries containing names of news channels, news anchors, and presidential candidates because they were too generic and not directly related to the audit topic. However, I kept the search queries where the names of the presidential candidates were together with election or election fraud related terms (e.g. 'Joe Biden voter fraud'). I also removed search queries that were in languages other than English. Finally, I had 88 search queries in total. Table 4.1 presents a sample.

4.2.2 Determining popular seed videos to collect up-next video trails

The second step of the audit experiment is to curate YouTube videos that would act as seed videos to collect the up-next video recommendation trails. I again leveraged Abilov et al's YouTube video dataset [37]. Recall, the authors identified clusters of Twitter users who either shared tweets promoting or detracting from voter fraud claims and released the YouTube videos related to election fraud 2020 shared by those users. At the the time of analysis, out of the ~12K videos present in the dataset,

8.9K were present on YouTube. The remaining videos were either removed or made private. Out of the videos that were still present, 1K videos were shared by users in the detractor cluster, 6.5K videos were shared by users in the promoting cluster, and the rest were shared by users who were suspended from Twitter. I sampled 445 videos that had accumulated the maximum number of views from both the promoting and detracting clusters (890 in total). Since the videos were not annotated by the authors for misinformation, I could not assume that videos shared by users in the promoting cluster would contain misinformation. Therefore, I conducted an intensive and iterative process to determine the labels and heuristics for annotating the YouTube videos for misinformation. I describe the process in detail in Section 4.2.6. Through the annotation process, I labeled the videos as supporting, neutral, or opposing election misinformation. Out of the 890 videos, 74 were opposing, 16 were neutral, 101 supported election misinformation while remaining were irrelevant. I selected the top 15 videos that had accumulated maximum engagement, determined by the number of views, for each stance (except the irrelevant) as seeds. Figure 4.3 illustrates the seed video curation method. Table 4.2 presents a sample of seed videos.

4.2.3 Experimental design

To conduct the crowd-sourced audit, I designed a Chrome browser extension named *TubeCapture* that enabled us to watch videos, conduct searches, and collect various YouTube components from users' browsers. Figure 4.4 presents an overview of the experimental design. To select the study participants, I conducted a screening survey of a large sample of people (details in Section 4.2.4). Next, participants were instructed on how to use *TubeCapture* and provided with a unique code to activate the extension. Once activated, they used *TubeCapture* for a period of 9 days. I seeded the extension with 45 seed videos and 88 search queries. For each participant, each day the extension opened YouTube in two browser windows, one standard window and one incognito window. While the personalized results act as treatment for the experiments, results obtained from incognito act as control since YouTube does not personalize content in the incognito browsing window [437]. By comparing the results from standard and incognito windows, I determine the role of YouTube's personalization algorithms in exposing users to misinformative content.

TubeCapture first collected and stored the user's YouTube homepage from standard and incognito windows. The extension ensured that the user had signed in to their YouTube account in the standard window and remained logged in using the same

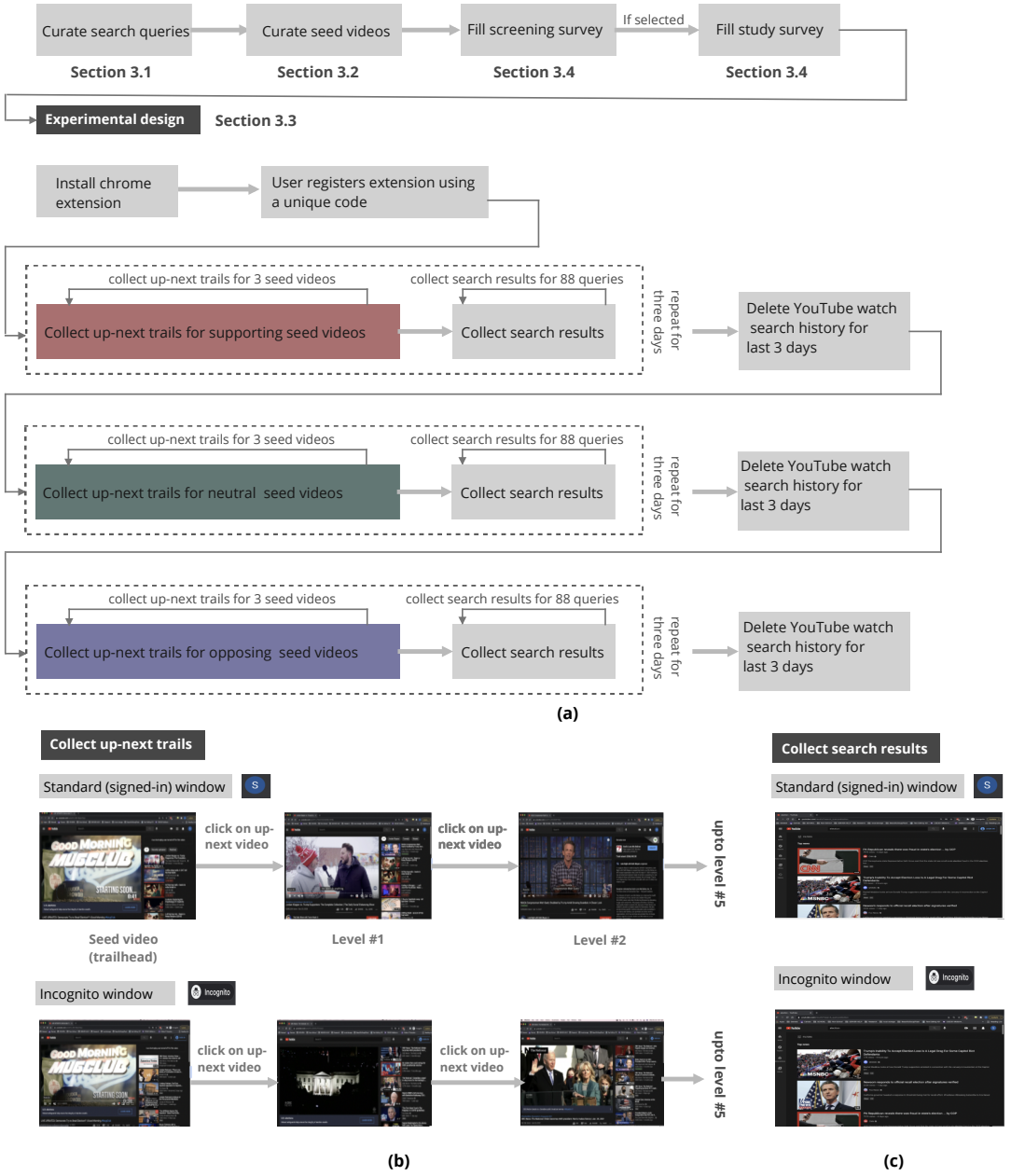


Figure 4.4: Figure (a) presents an overview of the crowd-sourced audit of YouTube for election misinformation, Figures (b) and (c) show how the extension *TubeCapture* collected YouTube components from both standard and incognito windows simultaneously.

YouTube account throughout the study period. I also ensured that the homepage from the standard window is stored without the user's email address to ensure the participant's anonymity. Next, the extension opened a seed video (previously selected) that supports election misinformation, watched it for 2 minutes, saved the video

page, clicked on the up-next video, and again saved the video page of the up-next video. This process was repeated until I collected 5 levels of up-next recommendations' video pages. I refer to the collection of 5 up-next video recommendations as up-next trails. Each day I collected up-next trails for five seed videos. Then, the extension again collected user's homepage followed by personalized (via standard window) and unpersonalized (via incognito window) search results for the curated search queries. The extension collected the search results for queries in the same order for every participant to control for carry-over effects of the search queries [187]. For days 1-3, the extension collected up-next trails for seed videos supporting election misinformation. At the beginning of the fourth day, the extension deleted the search and watch history created by the browser extension. According to YouTube, removing an item from search or watch history removes the impact of consuming that content on future searches and recommendations. This essential step helped in two ways- 1) it ensured that the history created by the extension in the first three days does not impact the rest of the experiment[3], and 2) it also ensured that the user histories built by the extension did not pollute users' future recommendations and search results after the study period is over. For days 4-6, the extension collected up-next trails for seed videos that were neutral in stance. At the beginning of the seventh day, again search and watch history developed by the extension was deleted. For days 7-9, the extension collected up-next trails for opposing seed videos. Towards the end of the 9th day, extension again deleted the YouTube history developed by the extension. All the data collected by the extension was sent to a back-end server. The participants were instructed on how to remove the extension after the study period was over. My current mixed design allows me to test how YouTube's algorithm fares under different conditions—watching videos of different stances—for individuals with different political beliefs. Note that I did not opt for a randomized assignment in a between-subject design since it would require a large number of participants to test all the conditions (3 political affiliations X 3 misinformation stances).

I built the YouTube capture extension using JavaScript libraries. The back-end server was set up using Flask and Nginx. I load-tested the server using Jmeter and ensured that the server could simultaneously handle 500 GET and 200 POST requests

[3]To test whether the Youtube algorithm discards the deleted history while making recommendations, the first author ran two test runs for two topics: presidential elections (the audit topic), and OLED TV. They built the account history of a brand new YouTube account for 3 days using a few videos and search queries related to the topic after which they deleted the search and watch history. Then the author manually inspected the homepage recommendations and the video recommendations of the top five videos present on the homepage and found that the effect of history deletion is almost immediate.

and added mechanisms to handle errors and server timeouts. I used a MySQL database for storing the data collected using the extension. The communication between the extension and the back-end server was encrypted using SSL. Note that to collect data, TubeCapture opened windows in the background of the currently active browser window, thereby allowing participants to continue working on their device while the extension is running. In case, the participant accidentally closed any of the windows opened by the extension, I informed users via a pop-up window and instructed them on how to resume running the extension.

After building the TubeCapture extension, I tested it within my research group and conducted three pilot studies. The aim of the pilot studies was to fix technical issues, examine the impact of running the extension on devices with different configurations, RAM, and operating systems as well as improve the usability of the extension.

4.2.4 Screening and study survey

In order to select participants for the study, I screened users according to several criteria. To be eligible for the study, users should be 1) 18 years of age or older, 2) reside in the United States, 3) have a YouTube account, 4) consume content on YouTube primarily in the English language, 5) have a chrome browser installed, 6) willing to run a chrome browser extension for 9 days and 7) have at least 8GB RAM on their device to ensure the smooth running of the extension[4]. The users who qualified for the screening survey were sent another study survey. The study survey contained questions about users' demographics, political affiliation, YouTube usage, trust in online information, their opinion on personalization and bias in various components of YouTube, and their view on the results of the presidential elections 2020 as well as conspiracies surrounding the elections. I also included two attention-check questions. The study survey was also used for screening participants. I disqualified users who 1) answered both attention check questions incorrectly, 2) did not frequently use YouTube, and 3) did not use YouTube to access news or information about the 2020 presidential elections. I also used the survey responses to obtain a balanced number of participants across three political affiliations (Democrats, Republicans, and Independents). Later in the recruitment phase, I had enough democrats and independents as participants and thus, added being a republican as a qualifying criterion in the study survey.

[4]I warned users against participating in the study if their device's RAM is less than 8GB and informed them that their device or browser might hang in such a situation

4.2.5 Recruitment and study deployment

For the pilot studies, I recruited users from a combination of platforms such as Reddit[5], Facebook ads, Twitter, and Amazon Mechanical Turk (AMT). The retention rate was highest for participants recruited from Twitter and AMT. Thus, I used these two platforms to recruit participants for the main study. Out of the 575 users who submitted the screening survey, 400 qualified, and 99 participated in the study. Out of the 99 participants, 94 ran the extension for the entire study duration. Overall, my study sample of 99 users constituted of 60.6% males and 39.39% females, was predominantly White/Caucasian (60.6%) and the majority (53.53%) of the participants had a bachelor's degree. Politically, 39.39% of the participants were Democrats, 34.34% independents, and 26.26% Republicans. Based on the results of 2020 presidential elections[6], 66.67% of the participants lived in the blue states, 32.32% in red while one individual resided in Puerto Rico[7].

4.2.6 Developing data annotation scheme

Developing the qualitative coding scheme to label YouTube videos for election misinformation was hard and time-consuming, requiring four rounds of discussions and consultation with an expert to reach a consensus on the annotation heuristics. In the first round, I and an undergraduate research assistant sampled 196 YouTube videos from Abilov et al's YouTube dataset [37] and separately annotated the videos. We considered prior work on election misinformation narratives [147] and YouTube content policy [436] as references to identify election misinformation, and came up with an initial annotation scale and heuristics to classify videos. Then we came together to reach a consensus on the annotation values. However, even after multiple rounds of discussions, annotations diverged for 33.6% of the videos. I then conducted additional rounds of annotation exercises with seven researchers, out of which five had extensive work experience on online misinformation. In every round, researchers independently annotated 15 videos and later discussed every video's annotation value and the researchers' annotation process. I also reached out to a postdoctoral researcher who has extensive research experience on online multi-modal election misinformation for feedback. Based on the insights provided by the external researchers and postdoc, I refined

[5]https://www.reddit.com/r/SampleSize/

[6]https://www.politico.com/2020-election/results/president/

[7]Puerto Rico is not considered a state but is considered an unincorporated territory of the United States

the annotation criteria and heuristics [8]. Below I describe the annotation guidelines and heuristics in detail.

4.2.6.1 Annotation guidelines

In order to annotate a YouTube video, the annotators were required to go through several fields present on the video page in the following order: title and description, the overall premise of the video which could be determined by going through the video transcript or watching the video content, and considering channel bias. I encouraged participants to perform an online search to gain more contextual information about events or individuals discussed in the video that they were unaware of. This strategy is grounded in lateral reading technique that is often used by fact-checkers for credibility assessments [424]. Note that I did not ask participants to consider video comments for the annotations because I found during the annotation exercises that comments could be misleading. For example, video *Dominion Voting Systems representative demonstrates voting machines* (Q7kPSzYsR6Y) contains a demonstration of dominion voting machines, however, the comments indicate the video to be supporting misinformation.

4.2.6.2 Annotation heuristics

In this section, I describe the annotation scale and heuristics.

Supporting election misinformation (1): This category includes YouTube videos that support or provide evidence for misleading narratives around the presidential elections. I did not include videos showing incidents of mail dumping, destroyed ballots, etc. in isolation. However, if the videos use these incidents to push a specific narrative/agenda like undermining confidence in mail-in voting, then I considered them as supporting misinformation. I also considered live YouTube videos (live press conferences, court hearings, etc.) that highlighted voter fraud claims without giving any additional context in the title, description, or beginning of the video as supporting misinformation. A few examples of videos in this category include *NO RETREAT! America Is About To #StopTheSteal | Good Morning #MugClub* (Xqcwzi8Onsk) where video's title, description, and content hint towards massive voter fraud incidents in the US 2020 presidential elections and *LIVE: Trump Legal Team Presents CLEAR Evidence of Fraud Before Georgia Senate Committee 12/3/20* (e35f4pUIYOg) which contains live

[8] It is important to note that all annotators and the post-doctoral researcher are left and center-left leaning individuals which may have affected how the content of YouTube videos was perceived and how the annotation heuristics were developed.

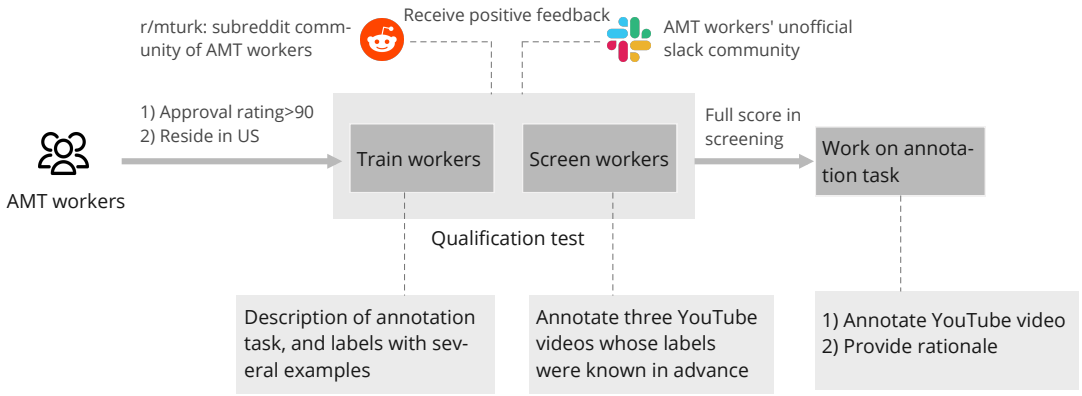


Figure 4.5: Figure illustrating the process of obtaining YouTube video annotations from AMT workers. The workers were screened via a qualification test where they were first trained by providing detailed descriptions of the annotation labels. To test their understanding, they were asked to annotate three YouTube videos whose labels were known in advance. Workers who correctly labeled the three videos proceeded to work on the annotation task. To ensure that the description of the annotation labels and task was clear and comprehensive, I posted on r/mturk—a subreddit community of AMT workers and AMT workers' unofficial slack channel. I released the qualification test and annotation task after receiving positive feedback from the AMT community.

footage capturing the testimony of individuals claiming occurrence of voter fraud in 2020 presidential elections. The video's description, title, and beginning do not contain any statements questioning or contradicting the claims of widespread voter fraud.

Neutral (0): I consider videos as neutral when they are related to the 2020 elections but do not support or oppose false narratives surrounding the elections. For example, video *WATCH: The first 2020 presidential debate* (w3KxBME7DpM) is considered neutral since it covers the first presidential debate of the elections.

Opposing (-1): I annotate videos as opposing when they oppose or debunk the misinformation narratives behind the 2020 US presidential elections. I also include satire videos making fun of the misinformative claims in this category. For example, video *Trump Has Yet To Show Real Evidence Of Fraud, But Getting Him Out Of Office May Be A Bumpy Ride* (7mJwuKhfvqY) whose title and description indicate that Donald Trump made false claims of massive voter fraud.

Other annotations: I mark a video as *Irrelevant* (2) if its content is not related to the presidential elections, as *URL not accessible* (3) if the YouTube video was not accessible at the time of annotation and as *Other languages* (4) when the content, title, or description of the YouTube video was in a language other than English.

4.2.7 Classifying YouTube videos for election misinformation

The crowd-sourced audit experiment resulted in ~47K unique YouTube videos and 35 unique YouTube shorts[9]. Given a large number of videos, I scaled the annotation process using a machine learning classifier. In this section, I present the method of creating the ground truth dataset, a description of features used in the classification model, model architecture, and the results of the classification.

4.2.7.1 Creating a ground truth dataset

Two researchers manually annotated 1196 videos using the guidelines and heuristics mentioned in Section 4.2.6. I obtained annotations for 545 additional videos using AMT. I describe the process of obtaining video annotations from AMT workers in Figure 4.5. Overall, in the ground truth dataset, I had 1741 videos out of which 124 are supporting[10], 257 opposing, 228 neutral, and 1132 irrelevant videos.

4.2.7.2 Feature description

I considered the following features for the classifier.
Snippet (title+description): I concatenated the title of the YouTube video with its description together, as done by [303], and used the concatenated string as a feature.
Transcript: Transcript contains the textual content of the video. I use transcripts auto-generated by YouTube.
Tags: Video tags are words that a content creator associates with their video while uploading it on the platform.
Video Statistics: Video statistics include the number of views, likes, comments, and date of publication.
Channel Bias: Since the election misinformation is closely entangled with the political beliefs [77, 264], I used partisan bias of YouTube channels as a feature. Using existing data sets on media bias and manual annotations (described in the next Section), I annotated YouTube channels' partisan bias on a 5-point scale of far-left to far-right.

Apart from the features listed above, I also tried several other features like LIWC dictionary [376], Credibility Cues [283], and hashtag matching from the Voter Fraud dataset on the text features [37] that didn't improve performance. Therefore, I do not discuss them in detail. Recall, while manually annotating the videos, I discovered that

[9]YouTube shorts are short YouTube videos with lengths equal to or less than 60 seconds
[10]Out of these 67 videos were removed from the platform at the time of analysis.

Classifier[Feature + Vectorizer + Imbalance Handling + Data]	Acc.	F1
SVM[Video Engagement Statistics]	0.38	0.14
SVM[Snippet + FastText]	0.61	0.56
SVM[Transcript + FastText]	0.58	0.51
SVM[Tags + FastText]	0.59	0.53
SVM[Snippet,Transcript,Tag + FastText]	0.63	0.57
SVM[Snippet,Transcript,Tag + Count]	0.65	0.58
SVM[Snippet,Transcript,Tag + TFIDF]	**0.71**	**0.65**
SVM[Snippet,Transcript,Tag,Channel Bias + Sentence Transformer]	0.73	0.69
SVM[Snippet,Transcript,Tag,Channel Bias + TFIDF]	0.74	0.70
SGD[Snippet,Transcript,Tag,Channel Bias + TFIDF]	0.64	0.57
KNN[Snippet,Transcript,Tag,Channel Bias + TFIDF]	0.61	0.58
XGB[Snippet,Transcript,Tag,Channel Bias + TFIDF]	0.74	0.68
Voting SVM + SGD + KNN + XGB [Snippet, Transcript, Tag, Channel Bias + TFIDF]	**0.75**	**0.71**
SVM[Snippet,Tag,Channel Bias + TFIDF + SMOTE + Additional Training Data]	**0.91**	0.90
XGB[Snippet,Tag,Channel Bias + TFIDF + SMOTE + Additional Training Data]	**0.91**	**0.91**

Table 4.3: A sample of of classifiers and feature set with the progression of performance.

comments are not a good indicator of the veracity of the video. Therefore, I chose not to include those in the feature set.

4.2.8 Annotating YouTube channels for partisan bias

The dataset of unique videos came from a large number of YouTube channels (~17.5K) and comprised of channels devoted to both news and non-news content. I coded the leaning of the channel on a 5-point Likert scale (far-left, center-left, neutral, center-right and far-right) using computational methods and several heuristics. First, to identify news-related channels, I used several pattern-matching techniques (e.g., finding keyword *news* in the channel's name, etc.) and discovered a total of 802 news channels. Then I used existing datasets on media bias from `mediabiasfactcheck.com` and `allsides.com` for annotating the channels. For channels whose annotations were not available in the datasets, I manually went through their title, description, sample videos, related information from their website, wikipedia, and/or google search to identify their leaning or the leaning of their affiliations. Many local news channels

such as KHOU[11] or KPRC[12] are affiliated with national channels. If I did not find the bias ratings for such local channels, I assigned them the label of their affiliations. For example, KHOU is associated with center-left CBS and thus, was also assigned a center-left rating. I assigned channels that didn't fall under news category the neutral label. I manually checked a random sample (n=50) of non-news channels and found only one channel that had content about news. Therefore, this process produced channel bias annotations (to be used as a feature in the classifier) with reasonable accuracy for the study, given that channel bias detection is not the main focus of the work.

4.2.8.1 Classifier Selection

To find a classifier that performs well on the dataset, I applied a series of machine learning classifiers on several combinations of feature sets. To create feature vectors, I tested two types of word vectors (count and tf-idf vectors) and two types of sentence vectors (FastText [13] and BERT [116]). For word vector generation, I cleaned the dataset by removing stop words and lemmatization, followed by up to 3-gram generation. To deal with data imbalance in the dataset, I used Synthetic Minority Over-sampling Technique [90] I applied several classifier models on the feature set including support vector machine, stochastic gradient descent, decision trees, nearest neighbor, and ensemble models. To find the best model, I performed a grid search on a five-fold cross-validation dataset by looking into standard parameter space for each classifier. For the sake of brevity, I only show a sample of combinations tested in Table 4.3. Out of all the combinations, both SVM and XGBoost performed the best (ACC=91%) when trained with snippet, tags, and channel bias features and tf-idf text vectorizer [14]. Based on Occam's Razor principle [103], I selected SVM as the final classifier, i.e., the simplest model with maximum accuracy. Using this classifier, I determined the annotation labels for the remaining videos. In total, the dataset consisted of 431 supporting, 1868 opposing, 1658 neutral, and 43041 irrelevant videos.

[11] https://www.youtube.com/c/KHOU

[12] https://www.youtube.com/c/KPRC2Click2Houston

[13] `https://fasttext.cc/`

[14] If I merge irrelevant and neutral videos into one class resulting in a three-class classification problem, SVM classifier performs with a 93% accuracy.

4.3 Ethical considerations

The browser extension *TubeCapture* uses crowd workers' YouTube accounts to watch videos (including videos containing election misinformation) and conduct searches on the platform. It was possible that participants would have seen more misinformation than they would have otherwise during and also after the research study due to the watch and search history built during the audit. In order to eliminate the potential harm of my experiments, I included two essential steps in the experimental design. First, the extension always opened the browser window in the background so that participants don't actively see the videos being played. Second, the extension deleted users' search and watch history built during the study period. Note that YouTube allows the deletion of items from the search and watch history for a specific date range. YouTube's website [404, 435] clearly states that "*search entries you delete will no longer influence your recommendations. At any time you can (also) remove videos (from watch history) to influence what YouTube recommends to you*". I explicitly informed users that their YouTube history during the study period would be deleted. I ensured that the extension expires after the study period so that it does not perform any action. In addition, I ensured that the YouTube pages saved by the extension do not contain users' personally identifiable information such as email addresses.

4.4 RQ1 Results: Extent of Personalization

To measure the extent of personalization in YouTube components, I compare the personalized list of video URLs present in the standard window with the baseline unpersonalized videos obtained from the incognito window. Below I discuss the metrics that I used to quantify personalization.

Measuring personalization in web search: In this study, to determine personalization in search results, I employ two metrics: jaccard index and rank bias overlap (RBO). Jaccard index measures the similarity between two lists and has been used in several previous audit studies to measure personalization in web search [187, 223, 236]. However, Jaccard index does not take into account the rank of the lists being compared. Thus, I used RBO metric introduced by Webber et al [413] which takes into account the order of elements in the list. The RBO function includes a parameter p which indicates the top-weightedness of the metric, i.e. how much will the metric penalize the difference in the top rankings. A previous audit study used the click-through rate

(CTR) of Google search results to estimate the value of p [331]. Because of the lack of CTR statistics available for YouTube, I consider the default value of p which is 1 (prior audit studies such as [249] opted for a similar approach), indicating that differences in all rankings are equally penalized. Both jaccard and RBO scores range between 0 and 1, with 1 indicating that the two lists have similar elements while 0 indicating that the lists are completely different.

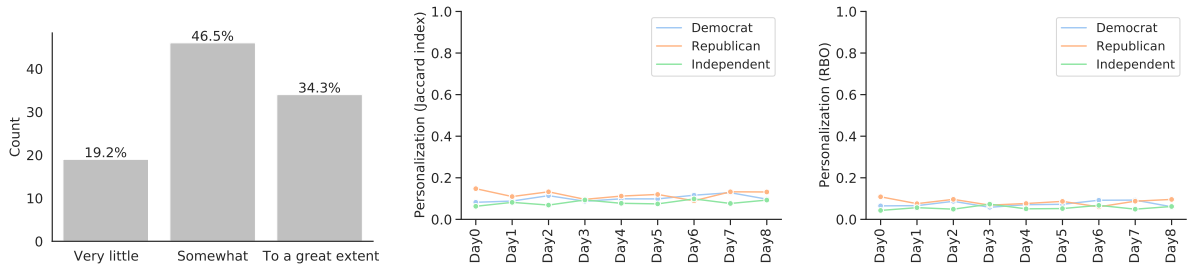
Measuring personalization in up-next trails: To measure personalization in up-next trails, I employ jaccard index and Damerau-Levenshtein (DL) distance [110]. DL distance is the enhanced version of edit distance that computes the number of transpositions in addition to insertions, deletions, and substitutions required to make the treatment list identical to the control list. DL distance has been used by prior audit work as a metric to estimate the ranking differences between two lists [81]. It returns a score from 0 to 1 (identical lists) indicating how similar the two lists are. I refrain from using the RBO metric to determine personalization in up-next trails because RBO is suitable for indefinite lists while the trails collected through the experiments have a known maximum length of five. I also refrain from using the Kendall tau metric since it requires the two ranked lists being compared to be conjoint[15]. Given, jaccard, RBO, and DL distance return similarity values, I define personalization as:-

$$1 - similarity\_metric(URL_{incognito}, URL_{standard}). \tag{4.1}$$

4.4.1 RQ1a: Personalization in search results

When asked in the study survey how much YouTube personalizes search results (Figure 4.6a), 34.34% believed YouTube personalizes search results to a great extent while 19.19% believed the extent of personalization to be very little. On quantitatively measuring the extent of personalization in YouTube search results, I found little to no personalization indicating that search results present in standard and incognito windows are highly similar. Figures 4.6b and 4.6c show the extent of personalization in SERPs calculated using jaccard index and RBO metric respectively for democrats, republicans, and independents for each day of the experiment run. I did not find any significant difference in the personalization values of SERPs for participants with respect to their political leaning.

[15]There are alternative versions of Kendall Tau that assume the dissimilar elements to be present at the end of the list. However, conceptually, the metric does not fit my collected trail data.



(a) Participant's belief in extent of personalization in YouTube search results

(b) Measuring extent of personalization in SERPs using jaccard index

(c) Measuring extent of personalization in SERPs using RBO

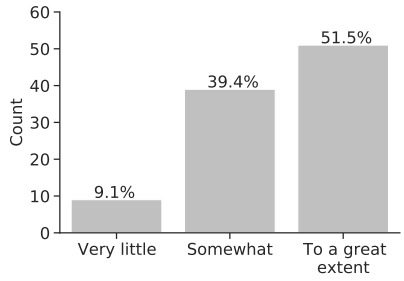
Figure 4.6: RQ1a results: Figure (a) shows participants' response to survey question: "How much, if at all, do you think YouTube personalizes search results". Figures (b) and (c) show personalization calculated via jaccard index values and RBO metric values respectively in YouTube's standard-incognito SERP pairs.

4.4.2 RQ1b: Personalization in up-next trails

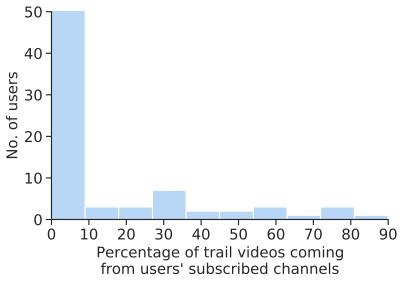
When asked how much YouTube personalizes up-next recommendations, 51.5% of participants believed that YouTube personalizes up-next recommendations to a great extent (refer Figure 4.7a). The quantitative measurements are in line with this belief showing that up-next trails are highly personalized. Figures 4.7c and 4.7d show the extent of personalization in up-next trails using jaccard index and DL distance. The graphs indicate that the up-next trails obtained from the users' standard and incognito windows are highly dissimilar and thus, highly personalized. Statistical test revealed that the amount of personalization in trails with supporting, neutral, and opposing seeds is significantly different [$F(2)=15.2$, $p<0.0001$]. Post hoc test revealed that up-next trails with seed videos opposing misinformation have lesser personalization (higher jaccard index[16]) when compared with up-next trails with supporting and neutral seed videos.

Next, I checked the influence of users' subscriptions on personalized trails. 81 (out of 99) participants had subscribed to at least one YouTube channel (mean=109.4, median=31, SD=207.8). The maximum number of subscriptions for a participant was 1073 and the minimum was 1. The participants had subscribed to 7670 unique channels out of which 79 either did not exist or were suspended due to violation of YouTube's moderation policy and thus, I did not consider these channels for analysis. To determine how many video recommendations in users' up-next trails were coming from their

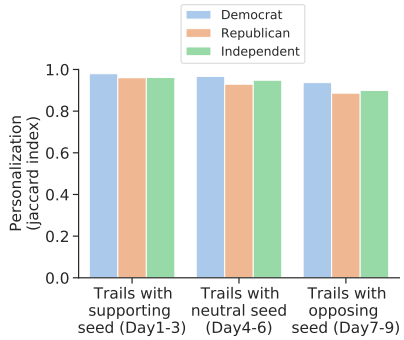
[16]The jaccard index values obtained were highly correlated with DL distance scores (pearson correlation coefficient = 0.96). Thus, I used jaccard index values to perform the statistical test.



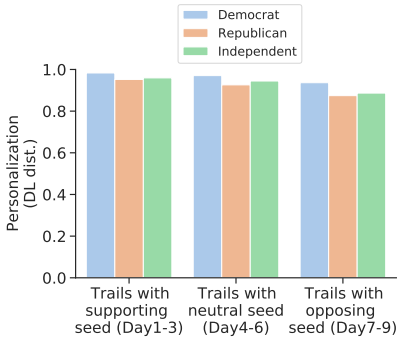
(a) Participant's belief in extent of personalization in YouTube up-next recommendations



(b) Distribution of percentage of up-next video recommendations coming from users' subscribed channels.



(c) Measuring extent of personalization using jaccard index



(d) Measuring extent of personalization using DL index

Figure 4.7: RQ1b results: Figure (a) shows participants' response to survey question: "How much, if at all, do you think YouTube personalizes up-next recommendations". Figure (b) shows the distribution of the percentage of YouTube videos recommended to the study participants from their subscribed channels. Figures (c) and (d) show personalization calculated via jaccard index values and DL distance metric values respectively in YouTube's standard-incognito up-next trails pairs.

subscriptions, first, for each user I extracted the unique videos recommended in all the up-next trails collected for the user. Then I filtered and calculated the number of videos coming from the users' subscribed channels. Figure 4.7b shows the distribution of the percentage of videos recommended to the participants in up-next trails that are coming from their subscribed channels. This percentage value is moderately correlated with the number of channels subscribed (r=0.61) and highly correlated with the number of news-related channels subscribed[17](r=0.71).

[17]To get a rough estimate of YouTube channels that broadcast news, I considered the news sources from `mediabiasfactcheck.com` and `allsides.com`. Additionally, I extracted the description of each channel and categorized it as a news channel if the description contained terms such as 'breaking news', 'politic*', 'current affairs', 'government', 'national tv', 'national news', 'international news',

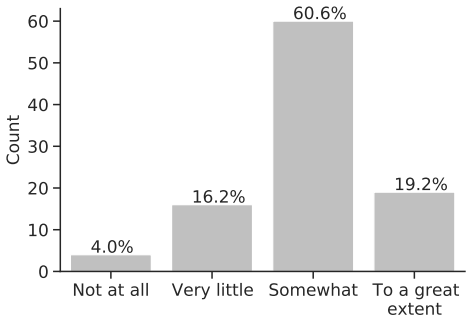
4.5 RQ2 Results: Amount of Misinformation

When asked how much do participants trust the credibility of videos in search results and recommendations, less than 20% reported that they trust the credibility of content shown to them by YouTube to a great extent (Figure 4.8). To determine how much credible information is presented by YouTube to users in reality, I quantify the misinformation present in the YouTube components by adopting the misinformation bias score developed in Chapter 3. The score determines the misinformation in ranked lists and is calculated as $\frac{\sum_{r=1}^{n}(x_r*(n-r+1))}{\frac{n*(n+1)}{2}}$; where x is the video annotation, r is rank of the video, and n is the total number of videos present in the SERP/up-next trail. To conform to the video annotation scale developed by me in Chapter 3, I map the annotation values to a normalized scale of -1, 0, and 1. I assign scores of -1 and 1 to videos opposing and supporting election misinformation respectively. Videos marked as irrelevant, neutral, belonging to a non-English language, or removed from the platform are assigned a 0 score. Thus, the misinformation bias score of a SERP/trail is a continuous value ranging between -1 (all videos are opposing election misinformation) to +1 (all videos are supporting election misinformation). Note that a positive score indicates a lean towards misinformation, while a negative score indicates a lean towards content opposing misinformation. For analysis, I consider the top ten search results and five consecutive videos in the up-next trails.

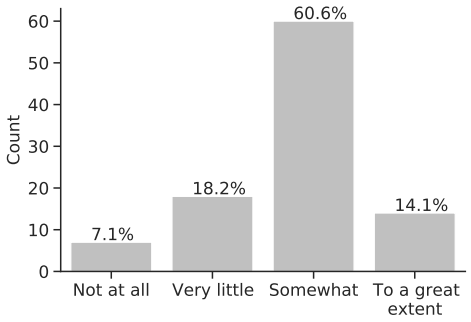
4.5.1 RQ2a: Misinformation in search results

The results of RQ1 showed that YouTube's SERPs are very slightly personalized suggesting that search results present in the standard and incognito windows are mostly similar. Therefore, to quantify the misinformation bias in SERPs I only consider the SERPs obtained from the standard YouTube windows of all the participants. I first calculated the average misinformation bias score for each of the 88 search queries for 9 days of the experiment run across all 99 participants. Figure 4.9 shows the distribution of misinformation bias scores for all the search queries. I observe that the average misinformation bias scores of 84 (out of 88) search queries are negative indicating that the search results contain more videos that oppose election misinformation as

'world news', 'global news', 'current affairs', 'wall street' etc. These terms were curated by the first author after manually going through the description of 50 national and regional news channels on YouTube. I found that 44 users had subscribed to news and politics-related channels.



(a) Participant's trust in the credibility of information presented in search results



(b) Participant's trust in the credibility of information presented in up-next recommendations

Figure 4.8: RQ2: Figure showing participants' response to survey question: "How much do you trust the credibility of information present in the " a) search results and b) up-next videos recommended by YouTube.

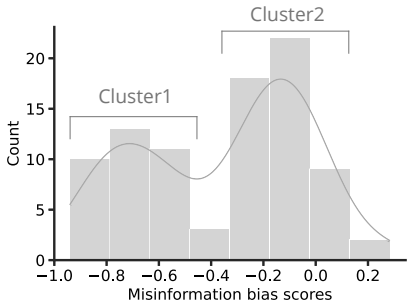


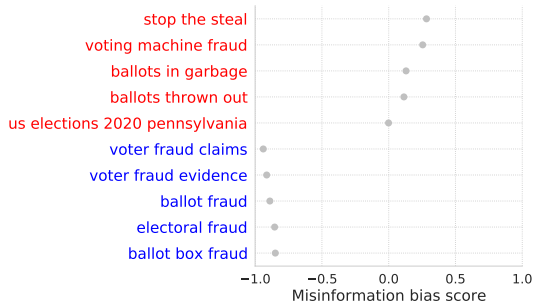
Figure 4.9: RQ2a results: Mean misinformation bias scores for 88 search queries for all participants. A negative score indicates that SERPs contain more videos opposing election misinformation.

Cluster1: Search queries containing keyword fraud in conjunction with keywords voter, election, and dominion
voter fraud evidence, dominion voter machine scandal, sharpie voter fraud, election fraud 2020, election fraud whistleblower
Cluster2: Search queries containing keywords election, and 2020
trump biden general election, presidential election 2020, presidential election results 2020, mail in ballots 2020

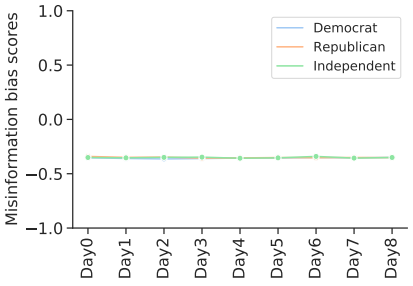
Table 4.4: The misinformation bias scores form a bimodal distribution, each constituting a cluster of similar queries. This table describes the clusters and presents sample queries for each cluster.

compared to videos supporting election misinformation[18]. Furthermore, I observe in Figure 4.9 that the misinformation bias scores of the SERPs form a bimodal distribution constituting two clusters of search queries (Table 4.4). The cluster1 search queries have the most negative bias, i.e. they contain more opposing videos. This cluster mostly

[18]Only four search queries in the query set ('stop the seal', 'voting machine fraud', 'ballots in garbage' and 'ballots thrown out') have a positive misinformation bias.



(a) Search queries with highest and lowest mean misinformation bias scores

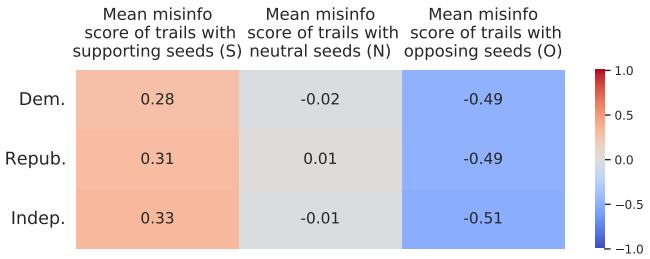


(b) Misinformation bias scores of search queries for each day of experiment run

Figure 4.10: RQ2a results: a) Search queries with highest (labeled in red) and lowest (labeled in blue) mean misinformation bias scores. Positive misinformation bias scores indicate a lean towards misinformation where as negative bias scores indicate a lean towards information that opposes misinformation. b) Figure showing the distribution of misinformation bias scores of search queries for democrats, republicans and independents. Note that the bias scores for the participants belonging to the different political leanings coincide indicating that misinformation bias in SERPs remain constant throughout for each participant.

consists of search queries containing the keyword *fraud* in conjunction with keywords *voter*, *election* and *dominion*. Cluster2 on the other hand consists of search queries with keywords *election* and *2020*. Overall, cluster1 consists of more search queries biased towards finding misinformation compared to search queries in cluster2. This indicates that YouTube pays more attention to search queries about election fraud and ensures that users are exposed to opposing videos when searching about fraudulent claims surrounding the elections.

Figure 4.10a shows five search queries with the highest and 5 search queries with the lowest misinformation bias. The search query 'voter fraud claims' has the least amount of misinformation bias, indicating that most of the search results for this query oppose election misinformation. On the other hand, the search query 'stop the seal' has the most amount of videos supporting election fraud claims. Next, I determine how do misinformation bias scores in SERPs vary for democrats, independents, and republicans. Figure 4.10b shows that the bias values for democrats, independents, and republicans for all days coincide indicating that the amount of misinformation bias is almost constant for all days for all participants irrespective of their partisanship. Overall, RQ2 results indicate that YouTube pushes debunking information in search results, more for search queries about voter fraud claims as compared to generic queries about the presidential elections.



heightPol. aff.	Statistical tests	Mean diff.
Democrats	F(2,3407)=4035.1 , p=0	S>N>O
Republicans	F(2,2265)=2981.4, p=0	S>N>O
Independents	F(2,2941)=3593.8, p=0	S>N>O

Figure 4.11: RQ2b results: Mean misinformation scores of standard up-next trails with seed videos that are supporting (S), neutral (N), or opposing election misinformation (O) for Democrats, Independents, and Republicans. A positive misinformation score indicates a lean toward misinformative content while a negative score indicates a lean toward content that opposes election misinformation. Statistical tests reveal a significant difference in the amount of misinformation contained in up-next trails. I find that democrats, republicans, and independents find more misinformation in supporting trails compared to neutral trails, and more misinformation in neutral trails as compared to opposing trails.

4.5.2 RQ2b: Misinformation in up-next trails

The results of RQ1 showed that participants' up-next trails are highly personalized. In other words, videos in up-next trails obtained from the standard window are different from videos in trails obtained from the incognito window. Recall, that trails extracted from incognito window act as baseline unpersonalized trails while trails extracted from the standard window, where users had signed into their accounts, act as personalized treatment trails. Therefore, to determine the impact of personalization on the amount of misinformation in up-next trails, I compare the misinformation bias scores of trails collected in standard windows with the trails collected in incognito windows. I find that the difference in misinformation bias scores of standard and incognito up-next trails is not significant (t=-0.62, p=0.53). This means that although the standard up-next trails are very different from the incognito up-next trails, there is no difference in the amount of misinformation present in them. To avoid inflating the sample size, for further downstream analysis, I only consider up-next trails obtained from participants' standard windows. This similar strategy was adopted by Robertson et al for analyzing bias in Google search results when they did not see any significant difference in the amount of partisan bias in incognito-standard SERP pairs [331].

4.5.2.1 Misinformation in standard up-next trails for different scenarios

In this section, I determine the amount of misinformation encountered by the study participants in the standard up-next trails for seed videos with different stances on elec-

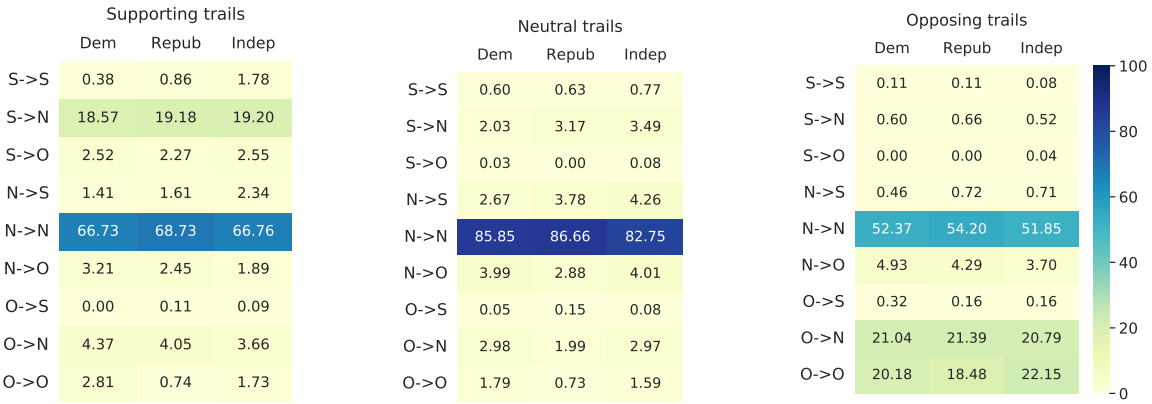
tion misinformation—supporting, neutral and opposing. Figure 4.11 shows the mean misinformation scores of different up-next trails collected from the standard windows of democrats, republicans, and independents. Recall that a positive misinformation score (>0) indicates a lean towards misinformation, while a negative misinformation score indicates a lean towards information that opposes election misinformation. I conduct within-group statistical tests to determine the difference in misinformation for the three scenarios (following trails for supporting, neutral, and opposing seed videos). The tests indicate a filter bubble effect. If users watch supporting videos, they are led to supporting videos in the trails. But if they watch neutral videos, they are led to less misinformation compared to when they watched supporting videos. However, if users watch opposing videos, they are led to more opposing videos in the up-next trails. The same trend is observed for democrats, republicans, and independents.

Is the amount of misinformation in trails with different seeds different for democrats, republicans, and independents? Between-group statistical tests reveal that amount of misinformation in supporting trails (KW H(2)=11.9,p=0.002) and neutral trails (KW H(2)=8.69,p=0.01) for democrats, independents, and republicans is significantly different. I find that independents receive more misinformation in their supporting trails as compared to democrats. Additionally, republicans receive more misinformation in their neutral trails compared to democrats.

Overall, by observing Figure 4.11, I realize misinformation scores of supporting trails are positive and opposing trails are negative. However, the magnitude of misinformation scores of opposing trails is much more than the supporting trails indicating that the strength of the filter bubble effect was more when the study participants watched videos opposing election misinformation.

4.5.2.2 Transitions in standard up-next trails

In this section, I gain more insights into the anatomy of YouTube's up-next trails by studying the various transitions present in them. This allows me to determine how users get pushed towards misinformative or debunking videos in the trails. Since the annotation scale consists of three values, supporting (S), neutral (N), and opposing (O), there are 9 transitions possible in the trails (S->S, S->N, S->O, N->S, N->N, N->O, O->S, O->N, N->O). For each participant, I first individually determine the percentage of each of these transitions present in the three types of standard up-next trails collected (ones starting with a supporting seed video, neutral seed videos, and opposing seed video). Then I calculated the mean percentage of all of



(a) Mean % of transitions in trails with seed videos supporting elec. misinfo.

(b) Mean % of transitions in trails with neutral seed videos

(c) Mean % of transitions in trails with seed videos opposing elec. misinfo.

Figure 4.12: RQ2b results: Mean percentage of various transitions present in the standard up-next trails of democrats, independents and republicans. S represents a video supporting election misinformation, N represents a neutral video and O represents a video opposing election misinformation. Transition S->S denotes that a YouTube video supporting election election misinformation leads to an up-next video recommendation supporting election misinformation.

these transitions for democrats, independents, and republicans. From Figure 4.12, I see that the maximum number of transitions across all participants and all types of up-next trails is N->N. Problematic transitions like S->S and O->S are less than 2% in trails of all users. However, comparatively S->S transitions are still more in the supporting up-next trails of independents (1.78%) compared to democrats (0.38%) and republicans (0.86%). In the neutral up-next trails of republicans and independents, N->S transitions dominate (after N->N transitions) indicating that independents and republicans are sometimes led to supporting videos in their up-next recommendations even when they are viewing neutral YouTube videos. I also observe that the opposing up-next trails majorly consist of transitions O->N and N->O (after N->N transitions) indicating that once a user watches a video that opposes election misinformation, YouTube pushes more videos that are either neutral or opposing in stance in the up-next trails of all the participants. I also observe that S->O transitions are less than S->N transitions in the supporting trails of democrats, republicans, and independents. Previous work has shown that watching YouTube videos that debunk misinformation helps in bursting filter bubbles of misinformation [390]. My work also shows that opposing videos could lead to more opposing videos (O->O transitions in opposing trails). Thus, increasing the number of S->O transitions can lead users to trustworthy

information on the platform.

4.5.3 RQ2c: Misinformation in homepages

I collected participants' YouTube homepages to determine how the bias in the homepage changes (δ) after watching a trail of videos starting with a seed video that is either supporting (δ_S), opposing (δ_O) or neutral (δ_N) in stance with respect to election misinformation. I calculated the impact of trails by using the following formula:-

$\delta_{stance} = Misinfo.\ score_{Homepage\_before\_the\_trail}$ - $Misinfo.\ score_{Homepage\_after\_the\_trail}$

δ_S, δ_N and δ_O represent the change in the amount of bias present in homepages because of watching a trail of up-next videos starting with supporting, opposing and neutral seeds. A negative δ would indicate that the YouTube homepage collected after the trail contained more opposing videos compared to the YouTube homepage before the trail. A positive δ, on the other hand, indicates either presence of more videos supporting election misinformation or a lesser number of opposing videos on the homepage collected after the trail as compared to the homepage collected before the trail. I consider the top ten recommendations present on the homepage for analysis. Figure 4.13 shows δ values for all three kinds of trails for democrats, republicans, and independents. I discuss a few results. I observe that after following the up-next video trails starting from a neutral seed, the homepages of democrats and independents contain more supporting videos. However, recall that the average misinformation score of the up-next trails with neutral seeds for both democrats and independents was negative (Figure 4.11). This indicates that although the up-next trails with neutral seeds lead users to more opposing videos, the homepages, however, contain more misinformation or a lesser number of opposing videos after the trail. I also observe that after watching up-next trail videos with supporting seed, republicans' homepage contain more opposing videos (Figure 4.13) while the trail itself contained more misinformation (Figure 4.11). However, note that the magnitude of the δ is low in all the conditions indicating that fewer videos supporting or opposing election misinformation appear on the participants' homepages.

4.6 RQ3: Composition and Diversity

In this research question, I want to characterize source diversity on YouTube when users search for election misinformation on the platform. Source diversity in searches

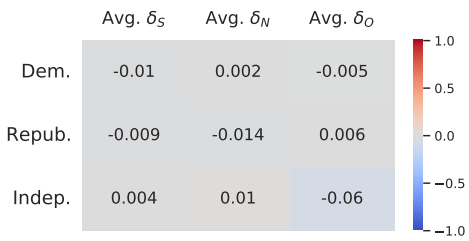


Figure 4.13: RQ2c results: Figure showing the average change in the amount of bias present in homepages because of watching a trail of up-next videos starting with supporting, opposing, and neutral seeds for democrats, republicans, and independents.

and recommendations is an important characterization of fairness [157]. Furthermore, given that the narratives about the election misinformation were closely intertwined with news sources and their leanings, it is important to determine what kinds of YouTube channels are users exposed to. News and media diversity can be characterized in multiple ways [221]. One typology characterizes media diversity with respect to *source* (content providers), *content* (perspectives) and *exposure* (actual consumption of diverse content) [288, 394]. My work analyzed the content diversity in RQ2 by analyzing the video's stance on election misinformation. I cannot study exposure diversity since it requires determining the actual content consumed (clicked, watched, etc) by the study participants in their naturalistic settings. For this study, I focus on source diversity in terms of the identity of top content providers (YouTube channels) and the distribution and concentration of channels in the standard SERPs and up-next trails. I acknowledge that future studies should also examine the ideological position of news sources and study the filter bubbles of partisan content on the platform.

4.6.1 RQ3a: Diversity in search results

For analysis, I consider the top ten search results in standard SERPs. Figure 4.14a shows the top 10 YouTube channels with impressions in the most number of search queries.[19] Here, I define impression as the occurrence of a channel's video in SERP. I observe that the left-leaning channel CNN on average appears in SERPs of more than half (61.86%) search queries. Additionally, except Fox news and 11Alive, all other top channels are left-leaning. I further analyzed which channels were responsible for the most relevant YouTube videos in the collected data. In the standard SERPs, I obtained a total of 4901 unique videos out of which 1940 (39.51%) videos were relevant, i.e.

[19]The top 10 YouTube channels and their mean percentage of total impressions were almost similar when calculated separately for democrats, republicans, and independents. Thus, I show the overall distribution for all users combined together.

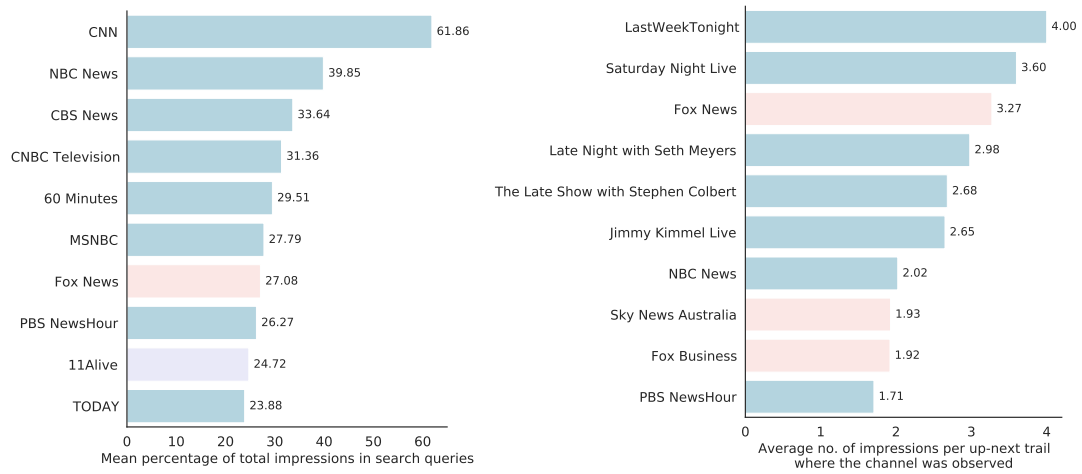


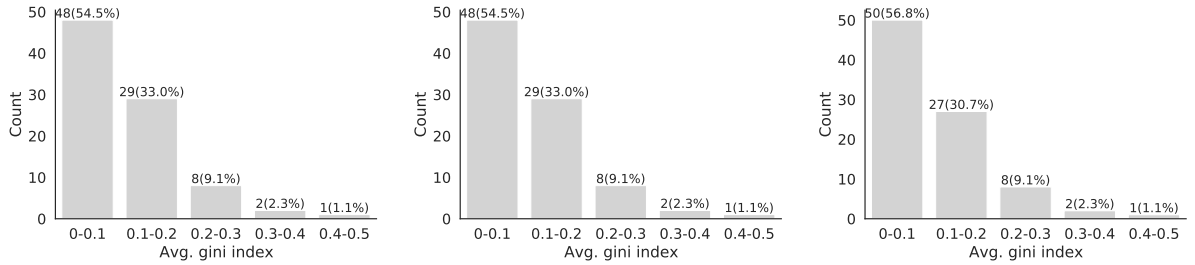
Figure 4.14: RQ3 results: a) Figure showing Top-10 YouTube channels with impressions in most number of search queries for all study participants. For example, on an average CNN appears in 61.86% of search queries for all study participants. b) Figure showing average number of impressions for Top-10 YouTube channels that appear in most number of standard up-trails collected for users. For example, on an average, videos from Fox News channel appear 3.27 times in those up-next trails where videos from the channel are observed. is a left-leaning channel, is right-leaning and is center-leaning.

related to elections (959 opposing, 865 neutral, and 103 supporting). Overall, in these relevant videos, most videos come from CNN and MSNBC. The most opposing videos come from channels MSNBC followed by CNN, most supporting videos come from Fox News followed by Daily Mail while most neutral videos come from NBC news followed by CNN. Given, CNN is one of the channels with the most opposing videos, it is encouraging to see that it has the most search query impressions.

Next, I determine the source diversity in the SERPs using gini coefficient metric [157, 394, 428]. Gini coefficient determines inequality in a frequency distribution. For the case of this study, I use this metric to determine the inequality in the distribution of YouTube channel impressions. For a given SERP consisting of videos from n unique channels, given a list of impressions for all YouTube channels [g_1, g_2,...g_n], then gini coefficient would be calculated as,

Gini coefficient (G) = $\frac{1}{2\bar{g}n^2} \sum_{i=1}^{|n|} \sum_{j=1}^{|n|} |g_i - g_j|$ where $\bar{g}$ is the mean of all impressions.

A fairer search engine would have lower values of gini coefficient indicating uniform distributions of YouTube channel impressions. Figure 4.15 shows the distribution of gini coefficients for all SERPs for democrats, republicans and independents. The



(a) Democrats **(b)** Republicans **(c)** Independents

Figure 4.15: RQ3a results: Distribution of Gini coefficients for all search queries (n=88) for a) Democrats, b) Republicans and c) Independents, calculated based on distribution of impressions of YouTube channels appearing in the search results.

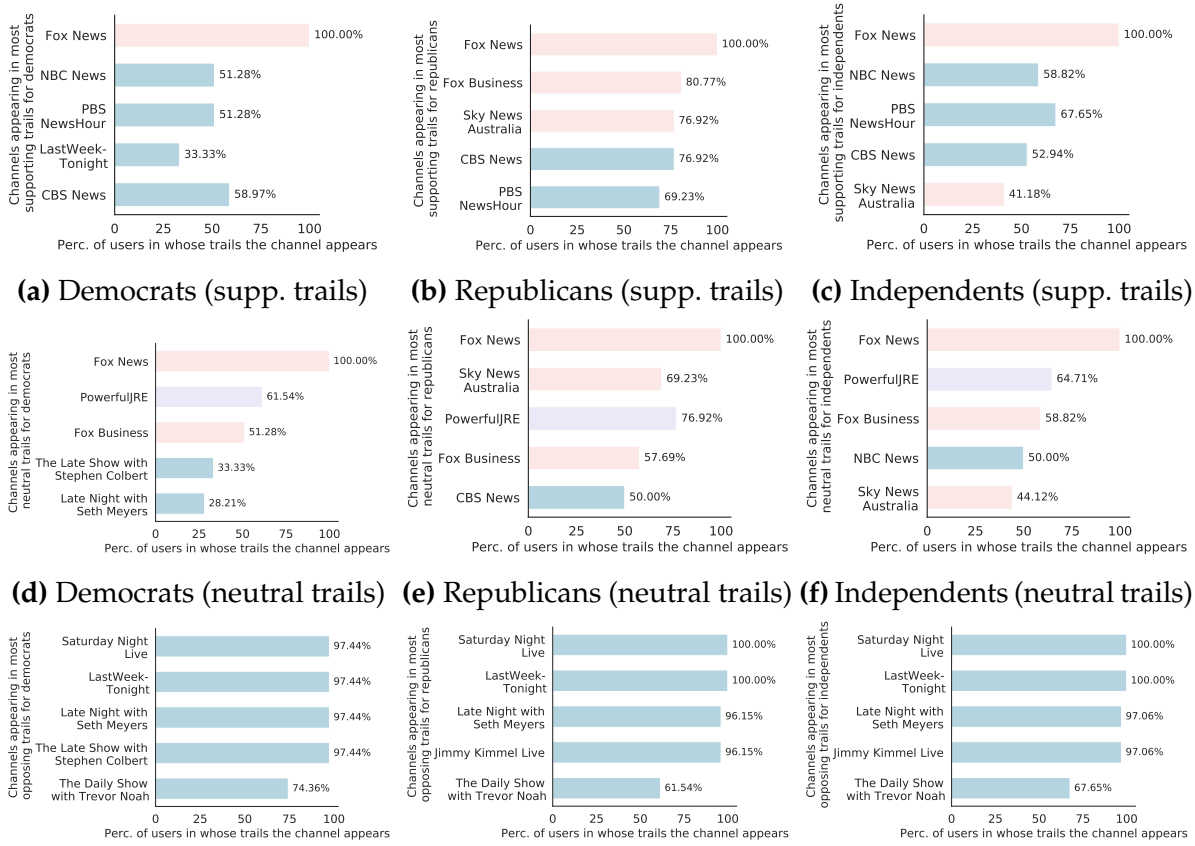
distributions are similar for users with different political leanings. Furthermore, for approximately 96% of search queries, the gini coefficient of SERPs is less than 0.3 indicating that YouTube has mostly evenly distributed videos from different channels in its search results.

4.6.2 RQ3b: Diversity in up-next trails

Overall, I collected 6943 videos in standard trails out of which 1082 are relevant, i.e. related to elections. The most number of opposing videos in trails come from channels MSNBC and Late Night with Seth Meyers*, most supporting videos in trails come from Fox News* and Fox Business, and most neutral videos come from Fox News* and NBC News[20]. Next, I determine the top ten YouTube channels occurring in the standard trails. Note, I do not consider the seed videos while analyzing the trails. Figure 4.14b shows the average number of impressions of the top 10 channels appearing the most number of times in the trails. Here, impression indicates the number of occurrences of a channel's videos in a trail, while considering trails containing videos from that channel. Note that the top channels are also channels of some of the seed videos in the dataset. The figure reveals that on an average, videos from LastWeekTonight, Saturday Night Live, and Fox News appear more than 3 times in a trail, when taking into account all the trails where the channel was observed. This finding indicates that videos from these channels lead to more videos from these channels in the up-next recommendations.

Next, to determine the diversity in trails, I determine the proportion of channels that are different than the channel of the seed video in the trails. I find that on average,

[20]* indicates that seed videos of the experiments also belonged to these channels.



(a) Democrats (supp. trails) **(b)** Republicans (supp. trails) **(c)** Independents (supp. trails)

(d) Democrats (neutral trails) **(e)** Republicans (neutral trails) **(f)** Independents (neutral trails)

(g) Democrats (oppos. trails) **(h)** Republicans (oppos. trails) **(i)** Independents (oppos. trails)

Figure 4.16: RQ3b results: Figure showing the top YouTube channels appearing in supporting, neutral, and opposing trails of democrats, republicans, and independents and the percentage of users in whose trails these channels appear. ■ is a left-leaning channel, ■ is right-leaning and ■ is center-leaning.

in an up-next trail of length five, I find 2.07 YouTube channels other than the channel of the seed video. The number of non-seed channels in up-next trails is least for trails with seed videos from Saturday Night Live (0.85), LastWeekTonight (0.86), and Late Night with Seth Meyers (1.07). Note, I did not calculate this metric for supporting, neutral, and opposing seeds separately since the channels of the supporting, opposing, and neutral videos are not unique. For example, I have a supporting as well as a neutral seed from Fox news. Given this scenario, there is no way to determine whether the videos appearing in the trails are due to the channel lean of the seed video or because of other factors. I also refrain from determining the diversity in up-next trails using gini coefficient since several trails had just one or two unique channels (M=3.1, SD=1.46) in which case gini coefficient would not give a good representation of diversity.

To get a sense of what kinds of channels are presented to users in the up-next

trails, I determine the channels appearing in the most number of trails of democrats, republicans, and independents for trails with supporting, neutral, and opposing seeds (Figure 4.16). I observe that Fox news appears in up-next trails with supporting and neutral seeds of all users. Fox Business and Sky News Australia appear in both the supporting and neutral up-next trails of more than half of the republicans (Figure 4.16b, and 4.16e). None of the seed videos belonged to these channels and they still appear in the up-next trails. Similarly, Sky News Australia also appears in the neutral up-next trails of 44.12% independents (Figure 4.16f) despite no neutral seed belonging to the channel. Furthermore, PowerfulJRE (Joe Rogan's YouTube channel) did not appear in the neutral up-next trails of all the users even though two neutral seed videos belonged to the channel (Figure 4.16d, 4.16e and 4.16f). On the other hand, the top channels appearing in the up-next trails with opposing seeds of all users (Figure 4.16g, 4.16h and 4.16i) are the channels of the opposing seed videos used in the experiment. Furthermore, three channels out of the top four appear in the trails of more than 96% of the users. This indicates that watching a video belonging to these left-leaning channels will probably lead to one or more videos belonging to this channel in the up-next recommendation trail.

4.7 Discussion

In this study, I conduct a crowd-sourced audit of the YouTube platform to determine how effectively the platform removed election misinformation from its various components. I discuss the implications of the findings below.

4.7.1 Standardization of search results

I find little to no personalization in the search results. I also did not find any effect of personalization on the amount of misinformation returned in search results. Throughout the study period, the amount of personalization and misinformation remained constant in the searches. On analyzing the standard SERPs, I find that YouTube returns more videos opposing election misinformation in 95% of the search queries that I tested. Interestingly, I see that misinformation scores of search queries having a misinformation lean (e.g. dominion voter fraud) are more negative compared to misinformation scores of queries that are neutral in stance (e.g. presidential election 2020). This finding implies that YouTube has paid more attention to the queries with

misinformation lean and ensured that users are exposed to more debunking information when they search about the fraudulent claims surrounding the elections. This selective attention is also in-line with results of past audits that showed YouTube improving the recommendations of topics like vaccination over 9/11 conspiracies [205].

My analysis also indicates that gini index of 96% of search queries is less than 0.3, with ~54% queries having a gini index of less than 0.1. Such low values of gini index imply that YouTube is ensuring source diversity in searches by evenly distributing videos from different channels in its SERPs. Furthermore, the distribution of gini coefficients was similar for all users irrespective of their partisanship. This finding indicates YouTube's attempt to expose users to videos from different channels rather than a select few based on participants' partisanship. Interestingly, in line with a previous audit on Google search [394], I find that CNN is one of the top channels whose videos appear in 61.8% of search queries. Future studies can test whether the dominance is due to emergent bias or the strategies adopted by the channel to enhance algorithmic visibility [394]. Overall, my analysis reveals that YouTube's search results are largely unpersonalized and the platform has had varying levels of success in removing misinformation and presenting videos that debunk election-related falsehoods in different clusters of search queries.

4.7.2 Scope for improvement in up-next trail recommendations

I find that up-next trails are highly personalized. However, for 50% of the users, only up to 10% videos in the up-next recommendations come from users' subscribed channels. Future audit studies should further investigate the impact of users' channel subscriptions (both news and non-news channels) on the platform's recommendations. I also find that there is no significant difference in the amount of misinformation that users are exposed to in up-next recommendation trails in the signed-in standard window and unpersonalized incognito window. On examining the standard up-next trails, I do find an echo-chamber effect. Users, irrespective of their partisanship, receive more misinformation in the up-next trails with supporting seeds as compared to the trails with neutral and opposing seeds (Figure 4.11). I also observe that the magnitude of misinformation scores of trails with opposing seeds is more than the magnitude of misinformation scores of trails with supporting seeds. This implies that users are exposed to a small number of misinformative videos when they follow the up-next recommendations of a video supporting election misinformation. On the other hand,

users are exposed to a larger number of opposing videos in the opposing up-next trails. This is a key finding also supported by prior work that showed that echo chambers of misinformation can be burst by watching debunking videos [390]. The platform can leverage this phenomenon by making its recommendation engine present more debunking videos to users which would then expose them to more credible videos in the recommendation trails.

I also examine various transitions in the up-next trails to study how users get pushed towards misinformation. Overall, I observe that problematic transitions where a supporting video is recommended in the up-next video recommendation of a supporting (S->S) or opposing video (O->S) are less than 2%. However, S->S transitions are more in trails with supporting seeds for independents compared to democrats and republicans. Furthermore, N->S transitions are also high in up-next trails with neutral seeds for independents. These findings are problematic. Showing misinformative videos to independents who might not have developed a strong opinion on the election fraud conspiracies could increase their chances of forming a pro-conspiracy belief. I also observe that N->S transitions are more for republicans in the up-next trails with neutral seeds (3.78%) compared to trails with supporting seeds (1.61%). This finding is again troublesome. Past studies have indicated that republicans are more susceptible to electoral fake news [300]. Thus, recommending videos supporting election misinformation to republicans watching neutral videos would expose them to more misinformation which might reinforce or lead to forming conspiratorial beliefs.

On analyzing the up-next trails for channel diversity, I observe several interesting phenomena. First, the number of impressions for left-leaning late-night show channels on YouTube such as LastWeekTonight is very high. On average, approximately 3-4 videos from these channels appear in the up-next trails (of length five) when starting with opposing seed videos. Furthermore, these channels appear in the video recommendations of almost all of the study participants. Similar to the late-night shows, I find that fox news also appears on average 3.27 times in the up-next trails of all participants. Future studies can look into the reasons behind the strong "algorithmic recognizability" [162] and high amplification of these channels in YouTube recommendations. Overall, I conclude that while YouTube has reduced misinformative videos in its up-next recommendations, there is still scope for improving the recommendation algorithm.

4.7.3 Participants' beliefs vs algorithmic reality

The study survey conducted before the audit experiment provided me with an opportunity to map participants' beliefs about personalization and trust in YouTube's algorithms with the reality of the situation as determined by the audits. The majority of participants believe that YouTube somewhat personalizes search results. However, in reality, they are hardly personalized. On the other hand, only half of the participants believe up-next recommendations to be highly personalized which is in line with the findings. This mismatch in beliefs and reality indicates users' lack of algorithmic awareness. It also acts as a call to action for the platform to make users aware of the functioning of the algorithms. Users could be made aware of personalization or lack of it by adding design features that promote algorithmic reflection, for example, seeing search results or recommendations of other users [69].

The survey also showed that, respectively, 19.2% and 14.1% users trust the credibility of information presented to them by YouTube in the search results and up-next recommendations to a great extent. This belief is problematic and indicates reliance on the platform's algorithms to show credible information. In reality, while I find the majority of YouTube's search results to be credible, up-next recommendations still contained misinformative videos. One way to make people spot misinformation on the platform and not blindly trust YouTube's recommendations could be by providing additional context about the content that the participant is searching for or viewing. While YouTube has started displaying Wikipedia links on the platforms [122], additional cues in the form of credibility citations, existing fact-checks or knowledge panel[21] could also be helpful [203].

4.8 Limitations and future work

My work is not without limitations. My audit study is observational in nature, i.e. my experiment does not isolate user attributes that produce the differences in misinformation measurements. I only make observations on the differences in misinformation received in searches and recommendations of users with different political affiliations. I recruited participants who used YouTube extensively to get information about the 2020 elections. However, for ethical reasons, I did not analyze participants' account histories to verify their self-reported data. My participant sample was also not balanced

[21] https://support.google.com/knowledgepanel/answer/9163198?hl=en

with respect to demographic attributes and political affiliation. I selected YouTube videos that had accumulated the most number of views as the seed videos for the audit experiments. One potential pitfall of such a sampling strategy is that it reduces the ecological validity of the experiment since the participants in the study might not have engaged with those videos in the past. Another limitation is that YouTube might have specifically tailored the recommendations of popular misinformative videos. Future studies could consider alternative strategies for sampling videos, such as selecting videos that were more recently published on YouTube or sampling a combination of videos that have accumulated the least and most amount of engagement. The search queries used in the audit also might not be representative of how the study participants formulate queries about the elections. Future studies can survey the study participants to determine how they used YouTube searches in the context of political elections as well as their information needs about the elections.

My classifier developed to annotate the YouTube videos for election misinformation has an error rate of 9% which could have affected the downstream analysis that I performed to quantify the amount of misinformation in various YouTube components. Additionally, I assign an annotation value of 0 to all videos that were removed from YouTube after the audit data collection. While the number of such videos is very small (<1%), it would result in a conservative estimate of misinformation bias present in the search results and recommendations. I use the misinformation bias score adopted from Hussein and Juneja et al's study that captures the amount of misinformation along with the rank of the video [205]. However, this metric does not take into account the relevance of the videos. Future studies can use metrics that measure simultaneously the relevance and credibility in ranked lists such as Normalised Weighted Cumulative Score and Convex Aggregating Measure [259]. In my audit experiment, after testing every condition (watching supporting, neutral, and opposing videos), I performed a step to delete users' YouTube history created by the extension so that it does not impact the other experimental condition. I tested out the effect of deletion on users' search and watch history for a few sample queries and videos and found that the effect of such deletion is almost immediate. However, I did not test out this scenario for all search queries and videos used in the audit. Future studies can determine how soon the deletion of history impacts users' recommendations and search results across various topics.

My study focuses on users' beliefs about the personalization and credibility of content on YouTube as well as the role of YouTube's algorithms in driving users

to the filter bubbles of problematic content. Future studies can focus on the impact of algorithmic recommendations on the radicalization of users. There are several scholars who argue that algorithms are not centrally culpable for the polarization or the filter bubbles that users experience on online platforms [75, 76, 417]. Many times the users of social media have a more diverse media diet than the non-users [75, 76]. Scholars posit that while algorithms can observe what a user consumes on social media, they cannot determine what the user actually prefers [108]. In other words, a digital choice is not always a true reflection of an individual's preference [108]. Furthermore, users might use different online platforms for different types of content [108]. Thus, to gain a holistic idea of the extent algorithms play a role in user polarization, future audit studies can conduct multi-platform crowd-sourced audits for individuals. These audit studies can determine the impact of algorithmic recommendations on users' social/political viewpoints via surveys and monitor users' patterns of content consumption simultaneously on multiple search engines and social media platforms used by the users.

4.9 Conclusion

In this study, I conducted a crowd-sourced audit on YouTube to determine the effectiveness of its content regulation policies with respect to election misinformation. I find that YouTube returns videos that debunk election misinformation in its searches. I also find that YouTube leads users to a small number of misinformative videos in up-next trails with seed videos that support election misinformation. Overall, my study shows that while YouTube has been largely successful in removing election misinformation from its searches, there is still scope to fix up-next recommendations.

CHAPTER 5

AUDITING E-COMMERCE PLATFORMS FOR HEALTH MISINFORMATION

5.1 Introduction

The period after the arrival of coronavirus witnessed an advent of dangerous health misinformation on the internet including anti-vaccine lies and a deluge of fraudulent treatments and cures [57, 143]. The pandemic also brought the focus back to the anti-vaccine movement which has gained popularity in the recent past with anti-vax social media accounts seeing 19% increase in their followers [301]. Health experts worry that vaccine hesitancy could make it difficult to achieve herd immunity against the Coronavirus. Battling health misinformation, especially anti-vaccine misinformation has never been more important.

Statistics show that people are increasingly depending on internet for health information [323] including information about medical treatments, immunizations, vaccinations and vaccine-related side effects [73, 148]. While internet search is convenient, relying too much on it for health information could be dangerous [229]. The algorithms powering the search engines are not traditionally designed to take into account the credibility and trustworthiness of the information. Thus, there has been a growing interest in empirically investigating the search engine results for vaccine misinformation. While multiple studies have performed audits on commercial search engines to investigate problematic behaviour [201, 205, 331] , e-commerce platforms

have received little to no attention ([95, 358] are two exceptions) despite critics calling platforms, like Amazon, a "dystopian" store for hosting several anti-vaccine books on its platform [121]. Amazon specifically has faced criticism from several technology critics for not regulating the content on its platform [65, 328]. Consider the most recent instance. Several medically unverified products for Coronavirus treatment like prayer healing, herbal treatments, antiviral vitamin supplements proliferated Amazon [127, 165], so much so that the company had to remove 1 million fake products from its platform after several instances of such treatments were reported by the media [143]. The scale of the problematic content on the platform suggests that Amazon could unintentionally be a great enabler of misinformation, especially health misinformation. It not only hosts problematic health-related content but its recommendation algorithms drive engagement by pushing potentially problematic content to users [164, 358]. Thus, in this study I investigate Amazon—world's leading e-retailer—for the most critical form of health misinformation i.e vaccine misinformation.

What is the amount of misinformation present in Amazon's search results and recommendations? How does personalization due to user history built progressively by performing real-world user actions, such as clicking or browsing certain products, impact the amount of misinformation returned in subsequent search results and recommendations? In this study, I dabble into these questions. I conduct 2 sets of systematic audit experiments– *Unpersonalized audit* and *Personalized audit*. In the *Unpersonalized audit*, I adopt Information Retrieval metrics from prior work [242] to determine the amount of health misinformation users are exposed to when searching for vaccine-related queries. In particular, I examine search-results of 48 search queries belonging to 10 popular vaccine-related topics like 'hpv vaccine', 'immunization', 'vaccination', 'MMR vaccine and autism', etc. I collect search results without logging in to Amazon to eliminate the influence of personalization. To gain in-depth insights about the platform's searching and sorting algorithm, the *Unpersonalized audits* ran for 15 consecutive days, sorting the search results across 5 different Amazon filters each day: "featured", "price low to high", "price high to low", "average customer review" and "newest arrivals," . The first audit resulted in 36,000 search results and 16,815 product page recommendations which were later annotated for their stance on health misinformation—promoting, neutral or debunking.

In the second set of audit—*Personalized audit*, I determine the impact of personalization due to user history on the amount of health misinformation returned in search results, recommendations and auto-complete suggestions. The user history is built

progressively over 7 days by performing several real-world actions such as "search" 🔍, "search + click" 🔍+👆, "search + click + add to cart" 🔍+👆+🛒, "search + click + mark top-rated all positive review as helpful" 🔍+👆+☑, "follow contributor" +👤 and "search on third party website" (Google.com in my case) . I collect several Amazon components in the *Personalized audit*, like homepages, product pages and pre-purchase pages, search results, etc. These components are explained in detail in Section 5.3.

I found Amazon hosting a plethora of health misinformative products belonging to several categories like Books, Kindle eBooks, Amazon Fashion (apparel, t-shirt, etc.) and Health & Personal care items (e.g. dietary supplements). Below I present the formal research questions, findings, contributions and implication of this study along with ethical implications.

5.1.1 Research Questions and Findings

In the first set of audit I ask,
RQ1 [*Unpersonalized audit*]: What is the amount of health misinformation returned in various Amazon components, given the components are not affected by user personalization?

> RQ1a: How much are search results contaminated with misinformation?
>
> RQ1b: How much are recommendations contaminated with misinformation? Is there a filter-bubble effect in the recommendations?

I find a higher percentage of products promoting health misinformation (10.47%) compared to products that debunk misinformation (8.99%) in the unpersonalized search results. I discover that Amazon returns high number of misinformative search results when users sort their searches by filter "featured" and high number of debunking results when they sort results by filter "newest arrivals". I also find Amazon ranking misinformative results higher than debunking results especially when results are sorted by filter "average customer reviews" and "price low to high". Overall, search results of topics "vaccination", "andrew wakefield" and "hpv vaccine" contain the highest misinformation bias when sorted by default filter "featured". My analyses of product page recommendation suggests that recommendations of products promoting health misinformation contain more health misinformation when compared to recommendations of neutral and debunking products. Next, in the second set of audit

I ask,

RQ2 [*Personalized audit*]: What is the effect of personalization due to user history on the amount of health misinformation returned in various Amazon components, where user history is built progressively by performing certain actions?

RQ2a: How are *search results* affected by various user actions?

RQ2b: How are *recommendations* affected by various user actions? Is there a filter-bubble effect in the recommendations?

RQ2c: How are *auto-complete suggestions* affected by various user actions?

The *Personalized audits* reveal that search results sorted by filters "average customer review", "price high to low", "price low to high" and "newest arrivals" along with auto-complete suggestions are not personalized. Additionally, I find that user actions involving clicking a search product leads to personalized homepages. I found evidence of filter-bubble effect in various recommendations found in homepages, product and pre-purchase pages. Surprisingly, the amount of misinformation present in homepages of accounts building their history by performing actions "search + click" and "mark top-rated all positive review as helpful" on misinformative product was more than the amount of misinformation present in homepages of accounts that added the same misinformative product in cart. The finding suggests that Amazon nudges users more towards misinformation once a user shows interest in a misinformative product by clicking on it but hasn't shown any intention of purchasing it.

Overall, the study suggests that Amazon has a severe vaccine/health misinformation problem exacerbated by its search and recommendation algorithms. Yet, the platform has not taken any steps to address this issue.

5.1.2 Contributions and Implications

In the absence of an online regulatory body monitoring the quality of content created, sold and shared, vaccine misinformation is rampant on online platforms. Through the work I specifically bring the focus on e-commerce platforms since they have the power to influence browsing as well as buying habits of millions of people. I believe the study is the first large-scale systematic audit of an e-commerce platform that investigates the role of its algorithms in surfacing and amplifying vaccine misinformation. My work provides an elaborate understanding of how Amazon's algorithm is introducing

misinformation bias in product selection stage and ranking of search results across 5 Amazon filters for 10 impactful vaccine-related topics. I found that even use of different search filters on Amazon can dictate what kind of content a user can be exposed to. For example, use of default filter "featured" lead users to more health misinformation while sorting search results by filter "newest arrivals" lead users to products debunking health-related misinformation. This is also the first study to empirically establish how certain real-world actions on health misinformative products on Amazon could drive users into problematic echo chambers of health misinformation. Both the audit experiments resulted in a dataset of 4,997 unique Amazon products distributed across 48 search queries, 5 search filters, 15 recommendation types, and 6 user actions, conducted over 22 (15+7) days [1]. The findings suggest that traditional recommendation algorithms should not be blindly applied to all topics equally. There is an urgent need for Amazon to treat vaccine related searches as searches of higher importance and ensure higher quality content for them. Finally, the findings also have several design implications that I discuss in detail Section 5.7.4.

5.1.3 Ethical Considerations

I took several steps to minimize the potential harm of my experiments to retailers. For example, buying and later returning an Amazon product for the purpose of the project can be deemed unethical and thus, I avoid performing this activity. Similarly, writing a fake positive review about an Amazon product containing misinformation could negatively influence the audience. Therefore, in the *Personalized audit* I explored other alternatives that could mimic similar if not the same influence as the aforementioned activities. For example, instead of buying a product, I performed "add to cart" action that shows users' intent to purchase a product. Instead of writing positive reviews for products, I marked top rated positive review as helpful. Since, accounts did not have any purchase history, marking a review helpful did not increase the "Helpful" count for that review. Through this activity, the account shows positive reaction towards the product, at the same time avoids manipulation and thus, eliminates impacting potential buyers/users. Lastly, I refrained from performing the experiments on real-world users. Performing actions on misinformative products could contaminate users' searches and recommendations. It could potentially have long-term consequences in terms of what types of products are pushed at participants. Thus, in the audit

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

experiments, accounts were managed by bots that emulated the actions of actual users.

5.2 Related work

5.2.1 Health misinformation in online algorithmic systems

The current research on online health misinformation including vaccine misinformation spans three broad themes: 1) quantifying the characteristics of anti-vaccine discourse [105, 280, 284], 2) building machine learning models to identify users engaging with health misinformation or instances of health misinformation itself [109, 158, 159] and 3) designing and evaluating effective interventions to ensure that users critically think when presented with health (mis)information [234, 399]. The existing research has a major gap. Most of these studies are post-hoc investigations of health misinformation, i.e the misinformation has already propagated and is in the wild now. The current work neither takes into consideration how the user encountered the misinformation nor does it investigate the role of the source of the misinformation. With the outset of internet, search engines have become the primary sources of information with 55% of American adults relying on the web to get medical information [323]. 5.9M people said that web search results influenced their decision to visit a doctor and 14.7M claimed that online information affected their decision on how to treat a disease [323]. Given how the medical information can directly influence one's health and well-being, relying on internet is not always a good idea. A lot of outlets have emerged that have contaminated the online health information. These sources could be conspiracy groups or websites spreading misinformation due to vested interests or companies having commercial interests in selling herbal cures or fictitious medical treatments [350]. Moreover, online curation algorithms themselves are not built to take into account the credibility of information. Thus, it is of paramount importance that role of search engines are investigated for harvesting health misinformation. How can I empirically and systematically probe the search engines to investigate problematic behavior like prevalence of health misinformation? In the next section, I briefly describe a new emerging research field called "algorithmic auditing" which is focused on investigating search engines to reveal problematic biases. I discuss this field as well as my contribution to this growing research in the next section.

5.2.2 Search engine audits

Search engines are modern day gatekeepers and curators of information, controlling "what" content users are exposed to. Their black-box algorithm can shape user behaviour, alter beliefs and even affect voting behaviour by impeding or facilitating the flow of certain kinds of information [117, 134, 238]. Despite their importance and the power they exert, till date, the search results and recommendations have mostly been unregulated. Information quality of search engine's output is still measured in terms of relevance and it is up to the user to determine the credibility of information. Thus, researchers have pushed for making algorithms more accountable. One recent method developed to achieve this is to perform systematic audits of search engines. Raji et al provide one definition of algorithmic audits. *An algorithmic audit involves the collection and analysis of outcomes from a fixed algorithm or defined model within a system. Through the stimulation of a mock user population, these audits can uncover problematic patterns in models of interest* [324].

Previous audit studies have investigated search engines for partisan bias [331], gender bias [93] and price discrimination [188]. However, only a few studies have systematically investigated the role of search engines in surfacing misinformation ([205] is the only exception). Moreover, there is a dearth of systematic audits focusing specifically on health misinformation. The past literature, mostly consists of small-scale experiments that probe the search engines with a handful of search queries. For example, an analysis of the first 30 pages of search results for query "vaccines autism" revealed that Google.com has 10% less anti-vaccine search results compared to other search engines, like Qwant, Swisscows and Bing [160]. Whereas, search results present in the first 102 pages for the query "autism vaccine" on Google's Turkey version returned 20% websites with incorrect information [136]. One recently published work, closely related to this work examined Amazon's first 10 pages of search results in response to query "vaccine". They only collected and annotated books appearing in the searches for misinformation [358]. The aforementioned studies probed the search engine for one single query and did the analysis on multiple search results pages. I, on the other hand, perform the *Unpersonalized audit* on a curated list of 48 search queries belonging to 10 most searched vaccine-related topics, spanning various combinations of search filters and recommendation types, over multiple days—an aspect missing in the previous work. Additionally I am the only ones who experimentally quantify the prevalence of misinformation in various search queries, topics and filters. Furthermore, instead of just focusing on books, I analyze the platform for products belonging to

different categories resulting in an extensive all categories inclusive coding scheme.

Another recent study on YouTube, audited the platform for various misinformative topics including vaccine controversies. The work established the effect of personalization due to watching videos on the amount of misinformation present in search results and recommendations on YouTube [205]. However, there are no studies investigating the impact of personalization on misinformation present in the product search engines of e-commerce platforms. My work fills this gap by conducting a second set of audit—*Personalized audit* where I shortlist several real-world user actions and investigate their role in amplifying misinformation in Amazon's searches and recommendations.

5.3 Amazon components and terminology

For the audit experiments, I collected three major Amazon components and numerous sub-components. I list them below.

1. **Search results:** These are products present on Amazon's Search Engine Results Page (SERP) returned in response to a search query. SERP results can be sorted using five filters: "featured", "price low to high," "price high to low," "average customer review" and "newest arrivals."

2. **Auto-complete suggestions:** These are the popular and trending search queries suggested by Amazon when a query is typed into the search box (see Figure 6.2c).

3. **Recommendations:** Amazon presents several recommendations to users as they navigate through the platform. For the purpose of this project, I collect recommendations present on three different Amazon pages: homepage, pre-purchase page and product pages. Each page is host of several types of recommendations. Table 5.1d shows the 15 recommendation types collected across 3 recommendation pages. I describe all three recommendations below.

    a) **Homepage recommendations:** These recommendations are present on the homepage of a user's Amazon account. The homepage recommendations could be of three types "Related to items you've viewed", "Inspired by your shopping trends" and "Recommended items other customers often buy again" (see Figure 5.1a). Any of the three types together or separately could be present on the homepage depending on the actions performed by the

user. For example, "Inspired by your shopping trends" recommendation type appears when a user performs one of two actions: either makes a purchase or adds a product to cart.

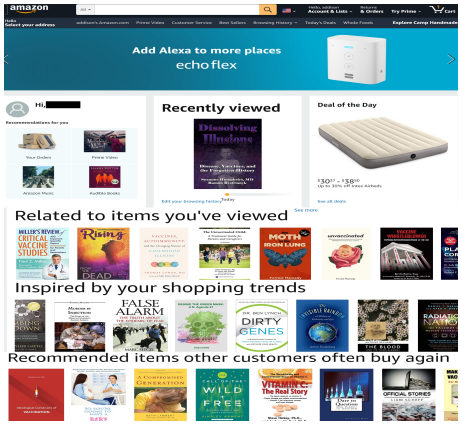
b) **Pre-purchase recommendations:** These recommendations consist of product suggestions that are presented to users after they add product(s) to cart. These recommendations could be considered as a nudge to purchase other similar products. Figure 5.1b displays pre-purchase page. The page has several recommendations like "Frequently bought together", "Customers also bought these highly rated items" and "Related to items you've viewed", etc. I collectively call these recommendations as pre-purchase recommendations.

c) **Product recommendations:** These are the recommendations present on the product page, also known as details page[2]. The page contains details of an Amazon product, like product title, category (e.g., Amazon Fashion, Books, Health & Personal care, etc.), description, price, star rating, number of reviews, and other metadata. The details page is home to several different types of recommendations. I extracted five: "Frequently bought together", "What other items customers buy after viewing this item", "Customers who viewed this item also viewed", "Sponsored products related to this item" and "Customers who bought this item also bought". Figure 5.1c presents an example of product page recommendations.

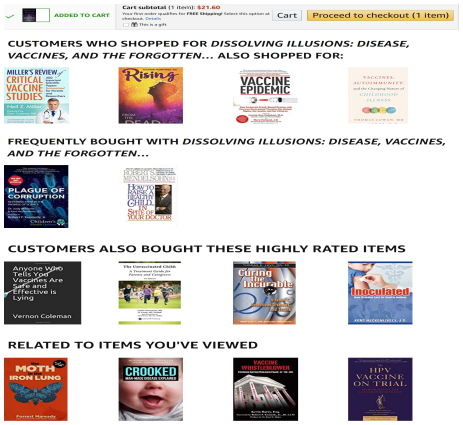
5.4 Methodology

Here I present the audit methodology in detail. This section is organized as follows. I start by describing the approach to compile high impact vaccine related topics and associated search queries. Then, I present overview of each audit experiment followed by the details of numerous methodological decisions I took while designing the audits. Next, I describe the qualitative coding scheme for annotating Amazon products for health misinformation. Finally, I discuss my approach to calculate misinformation bias in search results.

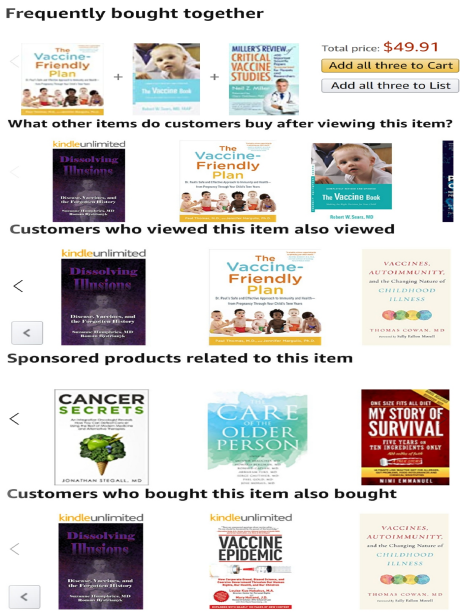
[2]https://sellercentral.amazon.com/gp/help/external/51



(a)



(b)

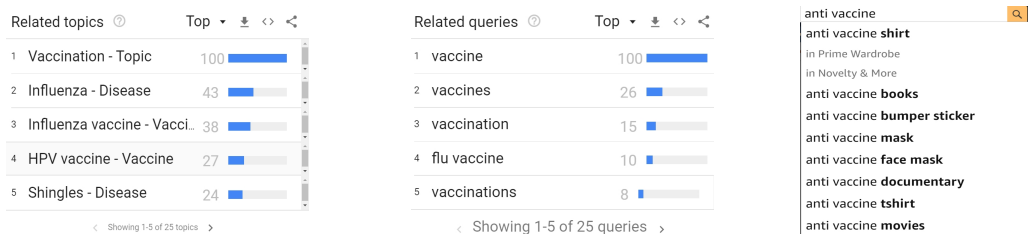


(c)

Recommendation page	Recommendation types
Homepage	Related to items you've viewed
	Inspired by your shopping trends"
	Recommended items other customers often buy again
Pre-purchase page	Customers also bought these highly rated items
	Customers also shopped these items
	Related to items you've viewed
	Frequently bought together
	Related to items
	Sponsored products related
	Top picks for
Product page	Frequently bought together
	Customers who bought this item also bought
	Customers who viewed this item also viewed
	Sponsored products related to this item
	What other items customers buy after viewing this item

(d)

Figure 5.1: (a) Amazon homepage recommendations. (b) Pre-purchase recommendations displayed to users after adding a product to cart. (c) Product page recommendations. (d) Table showing 15 recommendation types spread across 3 recommendation pages.



(a) **(b)** **(c)**

Figure 5.2: (a) Google Trend's Related Topics for *topic* vaccine. People who searched for vaccine topic also searched for these topics. (b) Google Trend's Related queries for *topic* vaccine. These are the top search queries searched by people related to vaccine topic. (c) Amazon's auto-complete suggestions displaying popular and trending search queries.

5.4.1 Compiling high impact vaccine-related topics and search queries

Here, I present my methodology to curate high impact vaccine-related topics and search queries.

5.4.1.1 Selecting high impact search topics:

The first step of any audit is to determine input—a viable set of topics and associated search queries that will be used to query the platform under investigation. I leveraged Google Trends (*Trends* henceforth) to select and expand vaccine-related search topics. *Trends* is an optimal choice since it shares past search trends and popular queries searched by people across the world. Since it is not practical to audit all topics present on *Trends*, I designed a method to curate a reasonable number of high impact topics and associated search queries, i.e., topics that were searched by a large number of people for the longest period of time. I started with two seed topics and employed a breadth-wise search to expand the topic list.

Trends allows to search for any subject matter either as a *topic* or a *term*. Intuitively, *topic* can be considered as a collection of terms that share a common concept. Searching as a *term* returns results that include terms present in the search query while searching as a *topic* returns all search terms having same meaning as the topic. For example, searching for "banana" as a *term* will return results that include terms like banana smoothie, banana, etc. On the other hand, searching for London as a *topic* will include results containing terms like Capital of UK, Londres (London in Spanish), etc [3]. I began

[3] `https://support.google.com/trends/answer/4359550?hl=en`

the search with two seed words namely "vaccine" and "vaccine controversies" and decided to search them as *topics*. Starting the topic search by the aforementioned seed topics ensured that the related topics will cover general vaccine-related topics as well as topics related to controversies surrounding the vaccines, offering us a holistic view of search interests. I set location to United States, date range to 2004-Present (this step was performed in Feb, 2020), categories to "All" and search service to "Web search". The date range ensured that the topics are perennial, and have been popular for a long time (note that *Trends* data is available from 1/1/2004 onwards). I selected the category setting as "All" so as to get a holistic view of the search trends encompassing all the categories together. Search service filter has options like was 'web search', 'YouTube search', 'Google Shopping', etc. Although Google shopping is an e-commerce platform like Amazon, its selection returned handful to no results. Thus, I opted for 'web search' service.

I employed *Trends's* Related Topics feature for breadth-wise expansion of search topics (see Figure 6.2a). I viewed the Related Topics using "Top" filter which presents popular search topics in the selected time range that are related to the topic searched. I manually went through the top 15 Related Topics and retained relevant topics using the following guidelines. All generic topics like Infant, Travel, Side-Effects, Pregnancy CVS, Virus, etc. were discarded. My focus was to only pick topics representing vaccine information. Thus, I discarded topics that were names of diseases but kept their corresponding vaccines. For example, I discarded topic Influenza but kept the topic Influenza vaccine. I also discarded temporal topics, such as 2009 flu pandemic vaccine. Moreover, 2009 flu pandemic was an Influenza pandemic and I had included influenza vaccine in my topic list [422]. I kept track of duplicates and discarded them from the search. To further expand the topics list, I again went through the Related Topics of the shortlisted topics and used the aforementioned filtering strategy to shortlist relevant topics. This step allowed me to expand the topic list to a reasonable number. After two levels of breadth-wise search, I obtained a list of 16 vaccine-related search topics (see Figure 5.3).

Next, I combined multiple similar topics into a single topic. The idea is to collect search queries for both topics separately and then combine them under one single topic. For example, topics zoster vaccine and varicella vaccine were combined since both the vaccines are used to prevent chickenpox. Therefore, search queries of both topics were later combined under topic varicella vaccine. All topics enclosed with similar colored boxes in Figure 5.3 were merged together. 11 topics remained after

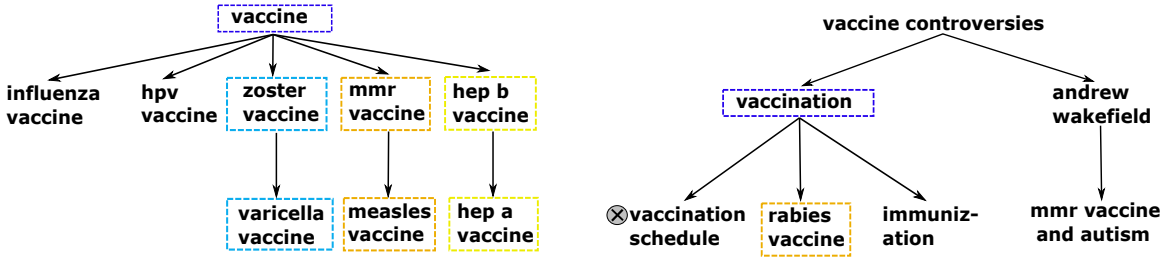


Figure 5.3: Figure illustrating the breadth-wise topic discovery approach used to collect vaccine-related topics from Google Trends starting from two seed topics: vaccine and vaccine controversies. Each node in the tree denotes a vaccine-related topic. An edge A→ B indicates that topic B was discovered from the Trends' Related Topic list of topic A. For example, topics "vaccination" and "andrew wakefield" were obtained from the Trend's Related Topic list of "vaccine controversies" topic. Then, topic "mmr vaccine and autism" was obtained from topic "andrew wakefield" and so on. ⊗ indicates the topic was discarded during filtering. Similar colored square brackets indicate similar topics that were merged together.

#	Search topic	Seed query	Sample search queries
1	vaccine controversies	vaccine controversy/ anti vaccine	anti vaccination anti vaccine shirt
2	vaccination	vaccine/ vaccination	vaccine vaccine friendly me
3	andrew wakefield	andrew wakefield	andrew wakefield wakefield autism
4	hpv vaccine	hpv vaccine	vaccine hpv hpv vaccine on trial
5	immunization	immunization	immunization immunization book
6	mmr vaccine and autism	mmr autism/ vaccine autism	autism autism vaccine
7	influenza vaccine	varicella vaccine	flu shot influenza vaccine
8	hepatitis vaccine	hepatitis vaccine	hepatitis b vaccine hepatitis a vaccine
9	varicella vaccine	varicella vaccine	chicken pox varicella vaccine
10	mmr vaccine	mmr vaccine	mmr vaccine measles vaccination

Table 5.1: Sample search queries for each of the ten vaccine-related search topics.

merging.

5.4.1.2 Selecting high impact search queries:

After shortlisting a reasonable number of topics, next I determine the associated search queries per topic, to be later used for querying Amazon's search engine. To compile search queries, I relied on both *Trends'* and Amazon's auto-complete suggestions; *Trends*, because it gives a list of popular queries that people searched on Google—the most popular search service, and Amazon, because it is the platform under investigation and it will provide popular trending queries specific to the platform.

Searching for a topic on *Trends* displays popular search queries related to the topic (see Figure 6.2b). I obtained top 3 queries per topic. Next, I collected Top 3 auto-complete suggestions obtained by typing seed query of each topic into Amazon's

search box (see Figure 6.2c). I remove all animal or pet related search queries (e.g "rabies vaccine for dogs"), overly specific queries (e.g. "callous disregard by andrew wakefield") and replaced redundant and similar queries with a single search query selected at random. For example search queries "flu shots" and "flu shot" were replaced with a single search query "flu shot". After these filtering steps, I had 48 search queries corresponding to 10 vaccine-related search topics. Table 5.1 presents sample search queries for all 10 search topics.

5.4.2 RQ1: Unpersonalized Audit

5.4.2.1 Overview

The aim of the *Unpersonalized audit* is to determine the amount of misinformation present in Amazon's search results and recommendations without the influence of personalization. I measure the amount of misinformation by determining the misinformation bias of the returned results. I explain the misinformation bias calculation in detail in Section 5.4.5. Intuitively, more the number of higher ranked misinformative results, higher the overall bias. I ran the *Unpersonalized audit* for 15 days, from 2 May, 2020 to 16 May, 2020. I took two important methodological decisions regarding which components to audit and what sources of noise to control for. I present these decisions as well as implementation details of the audit experiment below.

5.4.2.2 What components should we collect for the Unpersonalized audits?

I collected SERPs sorted by all five Amazon filters: "featured", "price low to high", "price high to low", "average customer review" and "newest arrivals". For analysis, I extracted the top 10 search results from each SERP. Recent statistics have shown that the first three search results receive 75% of all clicks [114]. Thus, I extracted the recommendations present on the product pages of the first three search results. I collected following 5 types of product page recommendations: "Frequently bought together", "What other items customers buy after viewing this item", "Customers who viewed this item also viewed", "Sponsored products related to this item" and "Customers who bought this item also bought". Refer Figure 5.1c for an example. I extracted the first product present in each recommendation type for analysis. Next, I annotated all collected components as promoting, neutral or debunking health misinformation. I describe the annotation scheme shortly in Section 5.4.4.

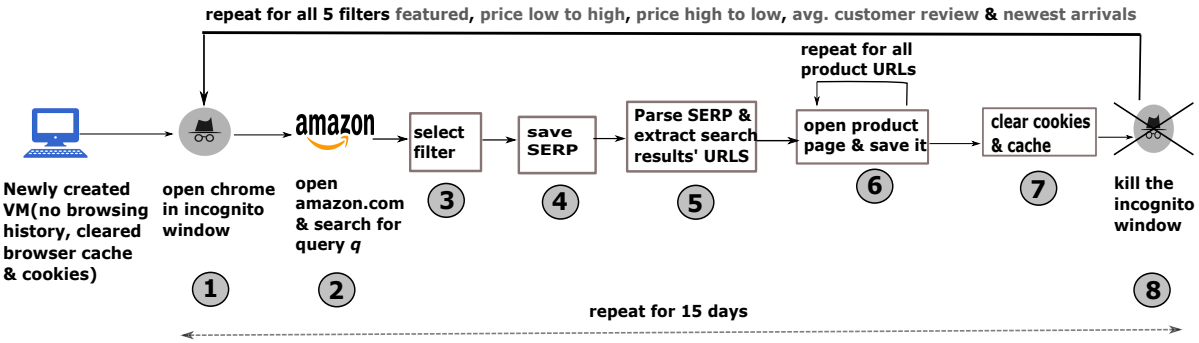


Figure 5.4: Eight steps performed in *Unpersonalized audit*. The steps are described in detail in Section 5.4.2.4

5.4.2.3 How can we control for noise?

I controlled for potential confounding factors that may add to noise to the audit measurements. To eliminate the effect of personalization, I ran the experiment on newly created virtual machines (VM) and freshly installed browser with empty browsing history, cookies and browser cache. Additionally, I ran search queries from the same version of Google Chrome in incognito mode to ensure that no history is built during the audit runs. To avoid cookie tracking, I erased cookies and cache before and after opening the incognito window and destroyed the window after each search. In sum, I performed searches on newly created incognito windows everyday. All VMs operated from same geolocation so that any effects due to location would affect all machines equally. To prevent machine speeds from affecting the experiment, all VMs had the same architecture and configuration. To control for temporal effect, I searched every single query at one particular time everyday for consecutive 15 days. Prior studies have established the presence of carry-over effect in search engines, where previously executed queries affect the results of the current query when both queries are issued subsequently within a small time interval [187]. Since, I destroyed browser windows and cleared session cookies and cache after every single search, carry over effect did not influence the experiment.

5.4.2.4 Implementation details

Figure 5.4 illustrates the eight steps for the *Unperonalized audits*. I used Amazon Web Services (AWS) infrastructure to create all VMs. I created selenium bots to automate web browser actions. As a first step, each day at a particular time, the bot opened

`amazon.com` in incognito window. Next, the bot searched for a single query, sorted the results by an Amazon filter and saved the SERPs. The bot then extracted the top 10 URLs of the products present in the results. The sixth step is an iterative step where the bot iteratively opened the product URLs and saved the product pages. In the last two steps, the bot cleared the browser cache and killed the browser window. I repeated steps 1 to 8 to collect search results sorted by all 5 Amazon filters ,'featured", "average customer review", "price low to high" and "newest arrivals". I added appropriate wait times after each step to prevent Amazon from detecting the account as a bot and blocking the experiment. I repeated these steps for 15 consecutive days for each of the 48 search queries. After completion of the experiment, I parsed the saved product pages to extract product metadata, like product category, contributors' names (author, editor, etc.), star rating and number of ratings. I extracted product page recommendations for the top 3 search results only.

5.4.3 RQ2: Personalized Audit

5.4.3.1 Overview

The goal of the Personalization Experiments is twofold. First, I assess whether user actions, such as clicking on a product, adding product to a cart would trigger personalization on Amazon. Second, and more importantly, I determine the impact of a user's account history on the amount of misinformation presented to them in their search results page, recommendations, and auto-complete suggestions; account history is built progressively by performing a particular action for seven consecutive days. I ran the *Personalized audit* from 12th August, 2020 to 18th August, 2020. I took several methodological decisions while designing this experimental setup. I discuss each of these decisions below.

5.4.3.2 What real-world user actions should we select to build account history?

Users' click history and purchase history trigger personalization and influence the price of commodities on e-commerce websites [188]. They also affect the amount of misinformation present in the personalized results [205]. Informed by the results of these studies, I selected six real-world user actions that could trigger personalization and thus, could potentially impact the amount of misinformation in search results and recommendations. The actions are (1) "search" 🔍 (2) "search + click" 🔍+👆 (3) "search + click + add to cart" 🔍+👆+🛒 (4) "search + click + mark top-rated all positive review

#	User action		Type of history	Tested values
1	Search product		Product search history	Product debunks vaccine or other health related misinformation (annotation value -1) Neutral health information (annotation value 0) Product promotes vaccine or other health related misinformation (annotation value 1)
2	Search + click product		Product search and click history	
3	Search + click + add to cart		Intent to purchase history	
4	Search + click + mark "Top rated, All positive review" helpful		Searching, clicking and marking reviews helpful history	
5	Following contributor by clicking follow button on contributor's page		Following history	
6	Search product on Google (third party application)		Third party search history	

Table 5.2: List of user actions employed to build account history. Every action and product type (misinformative, neutral or debunking) combination was performed on two accounts. One account sorted search results by filters "featured" and "average customer review". The other account built history in the same way but sorted the search results by filters "price low to high" and "newest arrivals". Overall, I created 40 Amazon accounts (6 actions X 3 tested values X 2 replicates for filters + 2 control accounts + 2 twin accounts).

as helpful" (5) "follow contributor" and (6) "search on third party website" (Google.com in my case). Table 5.2 provides an overview. First three actions involve searching for a product and/or clicking on it and/or adding it to cart. Through the fourth action, a user shows positive reaction towards a product by marking its top rated critical review as helpful. In the fifth action, a user follows a contributor. For example, for a product in the Books category, the associated list of contributors include the author and editor of the book. The contributors have dedicated profile pages that a user can follow. This action investigates the impact of following a contributor of a misinformative product. Sixth action investigates the effect of searching for an Amazon product on Google.com. The user logs into Google using the email id used to register the Amazon account. The hypothesis is that Amazon search results could be affected by third party browsing history. After selecting the actions, next I determine the products on which the actions needed to be performed.

#	Contributors to debunking health products		Contributors to neutral health products		Contributors to misinformative health products	
	name	url code	name	url code	name	url code
1	Paul-A-Offit	B001ILIGP6	Jason-Soft	B078HP6TBD	Andrew-J-Wakefield	B003JS8YQC
2	Seth-Mnookin	B001H6NG7A	Joy-For-All-Art	B07LDMJ1P4	Mary-Holland	B004MZW7HS
3	Michael-Fitzpatrick	B001H6L348	Peter-Pauper-Press	B00P7QR4RO	Kent-Heckenlively	B00J08DNE8
4	Ziegler-Prize	B00J8VZKBQ	Geraldine-Dawson	B00QIZY0MA	Jenny-McCarthy	B001IGJOUC
5	Ben-Goldacre	B002C1VRBQ	Tina-Payne-Bryson	B005O0PL3W	Forrest-Maready	B0741C9TKH
6	Jennifer-A-Reich	B001KDUUHY	Vassil-St-Georgiev	B001K8I8XC	Wendy-Lydall	B001K8LNVQ
7	Peter-J-Hotez	B001HPIC48	Bryan-Anderson	B087RL79G8	Neil-Z-Miller	B001JP7UW6

Table 5.3: List of contributors selected for building up account history for action "Follow contributors".

5.4.3.3 What products and contributors should we select for building account history?

To build user history, all user actions except "follow contributor" need to be performed on products. First, I annotated all products collected in the *Unpersonalized audit* run as debunking (-1), neutral (0) or promoting (1) health misinformation. I present the annotation details in Section 5.4.4. For each annotation value (-1, 0, 1), I selected top-rated products that had received maximum engagement and belonged to the most occurring category—'Books'. I started by filtering Books belonging to each annotation value and eliminated the ones that did not have an "Add to cart" button on their product page at the time of product selection. Next, I sorted the Books based on the accumulated engagement—number of customer ratings received by the Books. I again sorted the top 10 books obtained from the previous sorting based on star ratings received by the Books . I selected top 7 books from the second sorting for the experiment (see Table 5.4 for the shortlisted books).

Action "follow contributor" is the only action that is performed on contributors' Amazon profile pages [4]. I selected contributors who contributed to the most number of debunking (-1), neutral (0) and promoting (1) books. I retained only those who had a profile page on Amazon. Table 5.3 lists the selected contributors.

5.4.3.4 How do we design the experimental setup?

I performed all six actions explained in Section 5.4.3.2 and Table 5.2 on Books (or contributors of the books in case of action "follow contributor") that are either all debunking, neutral or promoting health misinformation. Each action and product type combination was acted upon by two treatment accounts. One account built its search history by first performing searches on Amazon and then viewing search results sorted

[4]The contributors could be authors, editors, people writing foreward of a book, publisher, etc.

#	Debunking products			Neutral products			Misinformative products		
	title (url code)	**S**	**R**	**title (url code)**	**S**	**R**	**title (url code)**	**S**	**R**
1	Vaccinated: One Man's Quest to Defeat the World's Deadliest Diseases (006122796X)	4.7	134	Baby's Book: The First Five Years (Woodland Friends) 144131976X	4.9	614	Dissolving Illusions: Disease, Vaccines, and The Forgotten History (1480216895)	4.9	953
2	Epidemiology and Prevention of Vaccine-Preventable Diseases, 13th Edition (990449114)	4.5	11	My Child's Health Record Keeper (Log Book) (1441313842)	4.8	983	The Vaccine Book: Making the Right Decision for Your Child (Sears Parenting Library) (0316180521)	4.8	1013
3	The Panic Virus: The True Story Behind the Vaccine-Autism Controversy (1439158657)	4.4	175	Ten Things Every Child with Autism Wishes You Knew, 3rd Edition: Revised and Updated paperback (1941765882)	4.8	792	The Vaccine-Friendly Plan: Dr. Paul's Safe and Effective Approach to Immunity and Health-from Pregnancy Through Your Child's Teen Years (1101884231)	4.8	877
4	Vaccines: Expert Consult - Online and Print (Vaccines (Plotkin)) (1455700908)	4.4	18	Baby 411: Your Baby, Birth to Age 1! Everything you wanted to know but were afraid to ask about your newborn: breastfeeding, weaning, calming a fussy baby, milestones and more! Your baby bible! (1889392618))	4.8	580	How to End the Autism Epidemic (1603588248)	4.8	717
5	Bad Science (865479186)	4.3	967	Uniquely Human: A Different Way of Seeing Autism (1476776245)	4.8	504	How to Raise a Healthy Child in Spite of Your Doctor: One of America's Leading Pediatricians Puts Parents Back in Control of Their Children's Health (0345342763)	4.8	598
6	Reasons to Vaccinate: Proof That Vaccines Save Lives (B086B8MM71)	4.3	232	The Whole-Brain Child: 12 Revolutionary Strategies to Nurture Your Child's Developing Mind (0553386697)	4.7	2347	Miller's Review of Critical Vaccine Studies: 400 Important Scientific Papers Summarized for Parents and Researchers (188121740X)	4.8	473
7	Deadly Choices: How the Anti-Vaccine Movement Threatens Us All (465057969)	4.2	223	We're Pregnant! The First Time Dad's Pregnancy Handbook (1939754682)	4.7	862	Herbal Antibiotics, 2nd Edition: Natural Alternatives for Treating Drug-resistant Bacteria (1603429875)	4.7	644

Table 5.4: Books corresponding to each annotation value shortlisted to build account histories in the *Personalized audit*. S represents the star rating of the product and R denotes the number of ratings received by the book.

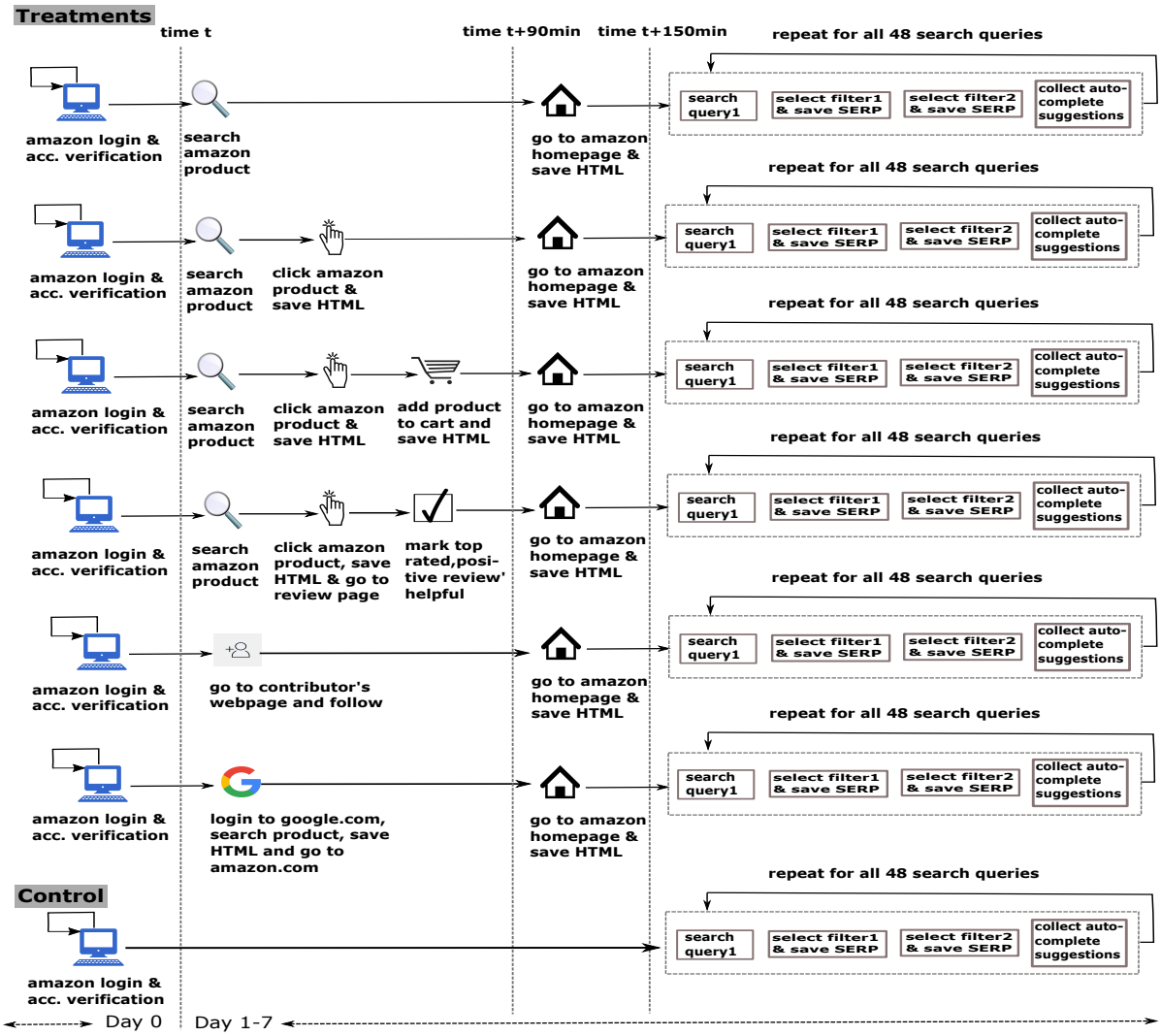


Figure 5.5: Steps performed by treatment and control accounts in *Personalized audit* corresponding to the 6 different features.

by filters "featured" and "average customer review" while the other did the same but sorted results by "price low to high" and "newest arrivals"[5]. I did not use the filter "price high to low" since intuitively it is less likely to be used during searches.

I also created 2 control accounts corresponding to 2 treatments that emulated the same actions as the treatment except that they did not build account histories by per-

[5]Every account created for this experiment was run by a bot. It was not possible for a bot to complete the following order of tasks in 24 hours because of wait times added between every subsequent actions–building history using a particular action, searching for 48 search queries sorted by 4 filters and collect auto-complete suggestions for those queries. Thus, every action-product type combination was performed on two accounts. First account, sorted the search results by two filters and second account sorted results using remaining two filters. I call these two accounts replicates since they built their history in the same way.

forming one of the 6 user actions. Like 2 treatment accounts, the first control account searched for 48 queries curated in Section 5.4.1.2 and sorted them by filters "Featured" and "Average customer Review" while the other control sorted them by the remaining two filters. Figure 5.5 outlines the experimental steps performed by treatment and control accounts. I also created twins for each of the control accounts. The twins performed the exact same tasks as the corresponding control. Any inconsistencies between a control account and its twin can be attributed to noise, and not personalization. Remember, Amazon's algorithms are a black box. Even after controlling for all known possible sources of noise, there could be some sources that I am not aware of or the algorithm itself could be injecting some noise in the results. If the difference between search results of control and treatment is greater than the baseline noise, only then it can be attributed to personalization. Prior audit work have also adopted the strategy of creating a control and its twin to differentiate between the effect due to noise versus personalization [188]. Overall, I created 40 Amazon accounts (6 actions X 3 tested values X 2 replicates for filters + 2 control accounts + 2 twin accounts). Next, I discuss the components collected from each account.

5.4.3.5 What components should we collect for the personalized audits?

I collected search results and auto-complete suggestions for all accounts and recommendations only for the treatment accounts. Search results were sorted by filters 'featured", "average customer review", "price low to high" and "newest arrivals". Once a user starts building their account history, Amazon displays several recommendations to drive engagement on the platform. I collected various types of recommendations spread across three recommendation pages—homepage, product pages and pre-purchase page. Pre-purchase pages were only collected for the accounts that perform "add to cart" action. Additionally, product pages were collected for accounts that clicked on search results while creating their respective account history. Each of the aforementioned pages consist of several recommendation types, such as "Customers who bought this item also bought", etc. I collected the first product present in each of these recommendation types from both product pages and pre-purchase pages and two products from each type from the homepages for further analysis. Refer to Table 5.1d and Figures 5.1a, 5.1b and 5.1c for examples of these recommendation types.

5.4.3.6 How do we control for noise?

To control for extraneous sources of noise, I adopted a number of measures from previous audit studies [187, 188]. First, all VMs had same configuration, architecture and operating system. Second, I ran all VMs from same geolocation to control for the effect of location. Third, I controlled for demographics by setting the same gender (Female) and Age (birth date 1/1/1995) for newly creating Google accounts. Recall, that these Google accounts were used to sign-up for the Amazon accounts. Since, the VMs were newly created, the browser had no search history that could otherwise hint towards users' demographics. Fourth, all accounts created their histories at the same time. They also performed the Amazon searches at the same time each day, thus, controlling for temporal effects. I also control for the category of the products used in building account histories. I selected books that have accumulated highest engagement for the experiments. Lastly, I did not account for carry over effects since it affected all the treatment and control accounts equally.

5.4.3.7 Implementation details

Figure 5.5 illustrates the experimental steps. I ran 40 selenium bots on 40 VMs. Each selenium bot operated on a single Amazon account. On day 0, I manually logged in to each of the accounts by entering login credentials and performing account verification. Next day, experiment began at time t. All bots controlling treatment accounts started performing various actions to build history. Note, everyday bots built history by performing actions on a single Book/contributor. At time t+90, bots collected and saved Amazon homepage. Later, all 40 accounts (control+treatment) searched for 48 queries with different search filters and saved the SERPs. Next, the bots collected and saved auto-complete suggestions for all 48 queries. I included appropriate wait times between every step to prevent accounts from being recognized as bots and getting banned in the process. I repeated these steps for a week. At the end of the week, for each treatment account I had collected personalized search results, recommendations and auto-complete suggestions. Next, I annotated the collected search results and recommendations to determine their stance on misinformation so that later I could analyze them to study the effect of user actions on the amount of misinformation presented to users in each component.

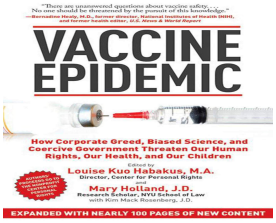
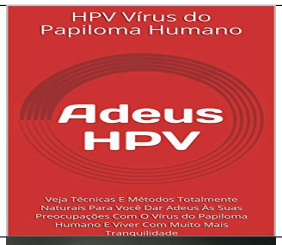
A. Scale Value	Annotation Description	Annotation Heuristics	Sample Amazon Products
-1	debunks vaccine misinformation	Product debunks, derides OR provides evidence against the myths/controversies surrounding vaccines OR helps understand anti-vaccination attitude OR promotes use of vaccination OR describes history of a disease and details how its vaccine was developed OR describes scientific facts about vaccines that help users to understand how they work OR debunks other health- related misinformation	
0	neutral health related information	All medicines and antibodies OR medical equipment (thermometer, syringes, record-books, etc.) OR dietary supplements that do not violate Amazon's policy OR products about animal vaccination and diseases OR health-related products not promoting any conspiratorial views about health and vaccines	
1	promotes vaccine and other health related misinformation	Product promotes disuse of vaccines OR promotes anti-vaccine myths, controversies or conspiracy theories surrounding the vaccines OR advocates alternatives to vaccines and/or western medicine (diets, pseudoscience methods like homeopathy, hypnosis, etc.) OR product is a misleading dietary supplement that violates Amazon's policy on dietary supplements- the supplement states that it can cure, mitigate, treat, OR prevent a disease in humans, but the claim is not approved by the FDA OR it promotes other health-related misinformation OR promotes other health-related misinformation	
2	unknown	Product's description and metadata is not sufficient to annotate it as promoting, debunking or neutral information	
3	removed	Product's URL is not accessible at the time of annotation	-
4	Other language	Product's title and description is in language other than english	
5	Unrelated	Non-health related products	

Table 5.5: Description of annotation scale, heuristics along with sample products corresponding to each annotation value.

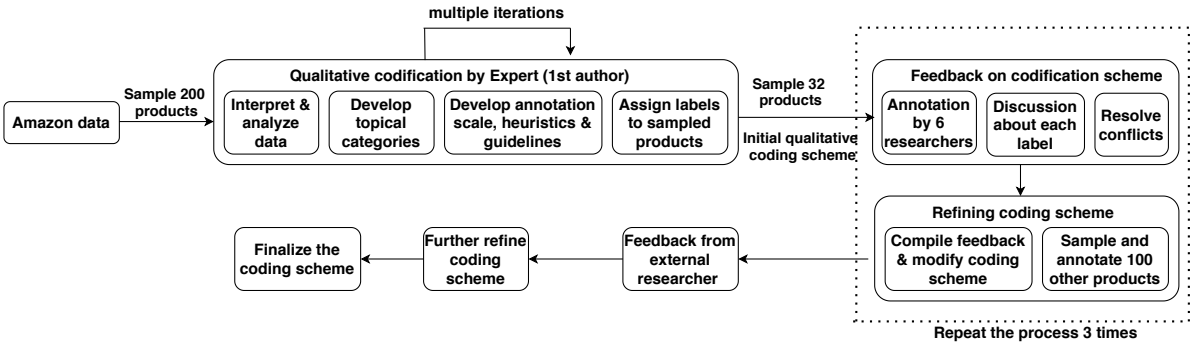


Figure 5.6: Qualitative Coding Process

5.4.4 Annotating Amazon data for health misinformation

Unlike determining partisan bias where bias could be determined by training models using features such as news source bias [331], labelling a product for misinformation is hard and time-consuming. There are no pre-determined sources of misinformation such as list of sellers or authors of misinformative products on Amazon. Additionally, I found that the annotation process for some categories of products, like Books, Kindle ebooks, etc. required me to consider the product image, read the book's preview, if available, and even perform external search about the authors. Therefore, I decided to manually annotate the data collection. I developed a qualitative coding scheme to label the Amazon data collection through an iterative process that required several rounds of discussions to reach an agreement on the annotation scale.

In the first round, the first author randomly sampled 200 Amazon products across different topics and categories. After multiple iterations of analyzing and interpreting each product, the author came up with an initial 7-point annotation scale. Then, six researchers with extensive work experience on online misinformation independently annotated 32 products, randomly selected from the 200 products. I discussed every product's annotation value and the researchers' annotation process. I refined the scale as well as the scheme based on the feedback. This process was repeated three times after which all six annotators reached a consensus on the annotation scheme and process. In the fourth round, I gathered additional feedback from an external researcher from the Credibility Coalition group[6]—an international organization of interdisciplinary researchers and practitioners dedicated to developing standards for news credibility and tackling the problem of online misinformation. The final result of the multi-stage iterative process (see Figure 5.6) is a 5-point annotation scale

[6]https://credibilitycoalition.org/

comprising of annotation values ranging from -1 to 3 (see Table 5.5). The scale provides an overview of the scientific quality of products users are exposed to when they make vaccine-related searches on Amazon.

5.4.4.1 Annotation Guidelines

In order to annotate the product, the annotators were required to go through several fields present on the product's detail page in the following order: title, description, top critical and top positive reviews about the product, other metadata present on the detail page, such as editorial reviews, legal disclaimers, etc. If the product is a book, the annotators are also recommended to do the following three steps: (1) go through the first few pages in the book preview [7] (2) see other books published by the authors, (3) perform a google search on the book and go through the first few links to discover more information about the book.

5.4.4.2 Annotation scale and heuristics:

Below I describe each value in my annotation scale.

Debunking (-1): Annotation value '-1' indicates that the product debunks vaccine misinformation or derides any vaccine-related myth or conspiracy theory or promotes the use of vaccination. As an example, consider the poster titled *Immunization Poster 1979 Vintage Star Wars C-3PO R2-D2 Original* (B00TFTS194)[8] that encourages parents to vaccinate their children. Products helping users understand anti-vaccination attitude are also included in this category. For example, consider a book titled *Health, Risk and News: The MMR Vaccine and the Media (Media and Culture)* (0820488380) which explores the controversy surrounding MMR vaccine and autism and investigates how media played a role in panicking the public. Moreover, products are also considered "debunking" if they describe the history about the development of vaccines or the science behind how vaccines work.

Promoting (1): Conversely, I annotated a product as '1' if it promotes any kind of vaccine or health-related misinformation. This category includes all products that

[7] Amazon has introduced a Look Inside feature that allows users to preview few pages from the book.

[8] Every title of the Amazon product is followed by a URL id. This URL id can be converted into a url using the format: http://www.amazon.com/dp/url_id

support or substantiate any vaccine related myth or controversies and encourages parents to raise a vaccine-free child. For example, consider the following books that promote anti-vaccination agenda. In *A Summary of the Proofs that Vaccination Does Not Prevent Small-pox but Really Increases It* (B01G5QWIFM), the author talks about dangers of large scale vaccination and in *Vaccine Epidemic: How Corporate Greed, Biased Science, and Coercive Government Threaten Our Human Rights, Our Health, and Our Children* (B00CWSONCE), the authors question vaccine safety and present several narratives of vaccine injuries. I annotated both books as 1. Several Amazon Fashion (B07R6PB2KP) products, Amazon Home (B01HXAB7TM) merchandise and cell phone accessories (B07Z9LDBD5) are also included in this category since they contained anti-vaccine slogans like "Educate before you Vaccinate", "Jesus wasn't vaccinated", etc.

I also include all products advocating any alternatives to vaccines in this category. Consider the book *Vaccine Free: Prevention and Treatment of Infectious Contagious Disease with Homeopathy* (1482789604) that not only encourages people to use homeopathy as a vaccine alternative to treat or prevent diseases but also instills fear in the minds of the public by discussing instances of vaccine injuries. Additionally, I include products that promote other health-related misinformation in this category. For example, the diet book titled *Natural Immune Support For People on the Go: Food, Diet, Immune Support Supplements, Colloidal Silver and Many Other Natural Remedies. Learn How to Naturally Boost Your Immune System!* (B086C1WT36) includes recipes with colloidal silver as an ingredient. According to the US Department of Health and Services, consumption of colloidal silver can be dangerous to health[9], and thus, this book was annotated with value '1'.

Dietary supplements that claim to cure diseases in their description but are not approved by Food and Drug Administration (FDA) are also included in this category.[10] For example, consider the dietary supplement *Yinchiao Tablet Herbal Supplement* (1586377791) that claims to cure pediatric ear infection, acute bronchitis, tonsillitis, pneumonia, pharyngitis, parotitis, measles, and influenza despite the claims not being approved by FDA. Not just dietary supplements, there are several books that claim to treat health conditions using unproven techniques. For example, the book *Weight Loss Hypnosis For Women: How to Lose Weight Quickly Using Meditation, Affirmations, And Other Hypnosis Techniques* found during the audit suggests self-hypnosis techniques to

[9] `https://www.nccih.nih.gov/health/colloidal-silver`

[10] Note that for the dietary supplements category, Amazon asks sellers not to state that the products cure, mitigate, treat, or prevent a disease in humans in their details page, unless that statement is approved by the FDA [87]

help lose weight (B0881V7RBL).

Neutral (-0): I annotated all medical equipment and medicines as neutral (annotation value '0'). Note that it is beyond the scope of this project to determine the safety and veracity of the claims of each medicine sold on the Amazon platform. This means that the number of products that I have determined to be promoting (1) serve as the lower bound of the amount of misinformation present on the platform. This category also includes dietary supplements that do not violate Amazon's policy and pet/animal-related products. Health-related products not advocating a conspiratorial view are also included in this category.

Other annotations: I annotated a product as '2' if the product's description and metadata were not sufficient to determine the stance of the product. I assigned values '3' and '4' to all products whose URL was not accessible at the time of the annotation and whose title and description was in a language other than English, respectively. I annotated 'all non-health related products (e.g. diary, carpet, electronic products, etc.) with value '5'. Table 5.5 presents examples of products belonging to these categories.

Both the audits resulted in a dataset of 4,997 Amazon products that were annotated by the first author and Amazon Mechanical Turk workers (MTurks). The first author being the expert annotated majority of products (3,367) to determine what would be a good task representation to obtain high quality annotations for the remaining 1,630 products from novice MTurks. I obtained three Turker ratings for each remaining product and used the majority response to assign the annotation value. My task design worked. For 97.9% of the products, annotation values converged. Only 34 products had diverging responses. The first author then annotated these 34 products to obtain the final set of annotation values. I describe the AMT job in detail in the next section.

5.4.4.3 Amazon Mechanical Turk Job

Turk job description: In this section, I describe how I obtained annotations for the study from Amazon Mechanical Turk workers (MTurks). Past research has shown that it is possible to get good data from crowd-sourcing platforms like Amazon Mechanical Turk (AMT) if the workers are screened and trained for the crowd-sourced task [282]. Below I describe the screening process and the annotation task briefly.

Screening: To get high quality annotations, I screened MTurks by adding 3 qualification requirements. First, I required MTurks to be Masters. Second, I required them

to have atleast 90% approval rating. And lastly, I required them to get a full score of 100 in a Qualification Test. I introduced a test to ensure that MTurks attempting the annotation job had a good understanding of the annotation scheme. The test had one eligibility question asking them to confirm whether they are affiliated to authors' University. Other three questions involved Mturks to annotate three Amazon products. First author annotated these products and thus, their annotation values were known. To ensure MTurks understood the task and annotation scheme, I gave detailed instructions and described each annotation value in detail with various examples of Amazon products in the qualifying test. Examples were added as visuals. In each example, I marked the meta data used for the annotation and explained why a particular annotation value was assigned to the product.

I took two steps to ensure that instructions and test questions were easy to understand and attempt. First, I posted the test on subreddit r/mturk[11]—a community of MTurks, to obtain feedback. Second, I did a pilot run by posting ten tasks along with the aforementioned screening requirements. After obtaining positive feedback from the community and successful pilot-run, I released the AMT job titled "Amazon product categorization task". I paid the Turks according to the United States federal minimum wage ($7.25/hr). Additionally, I did not disapprove any worker's responses.

Amazon product categorization task: I posted 1630 annotations (tasks) in batches of 50 at a time. The job was setup to get three responses for each annotation value. The majority response was selected to label the Amazon product. To avoid any MTurk bias, I did not explicitly reveal that the idea behind the task was to get misinformation annotations. I used the term "Amazon product categorization" to describe the project and task throughout. For 34 products, all three MTurk responses differed. The first author then annotated these products to get annotation values.

5.4.5 Quantifying misinformation bias in SERPs:

In this section, I describe the method to determine the amount of misinformation present in search results collected in both sets of audits. How do I estimate the misinformation bias present in Amazon's SERPs? First, I used the annotation scheme to assign misinformation bias scores (s_i) to individual products present in SERPs. I converted the 7 point (-1 to 5) scale to misinformation bias scores with values -1, 0 and 1. I mapped annotation values 2, 3, 4, and 5 to bias score 0. Because of the mapping, the

[11]https://www.reddit.com/r/mturk/

bias calculations will give a conservative estimate (lower bound) of misinformation bias present in the search results. Now, a product can be assigned one of the three bias scores: -1 suggests that product debunks misinformation, 0 indicates a neutral stance and 1 implies that the product promotes misinformation. Next, to quantify misinformation bias in Amazon's SERPs, I adopt the framework and metrics proposed in prior work to quantify partisan bias in Twitter search results [242]. Below I discuss three kinds of bias proposed by the framework and delineate how I estimate each bias with respect to misinformation. Table 5.6 illustrates how I calculated the bias values.

(i) The *input bias* (*ib*) of a list is the mean of misinformation bias scores of the constituting elements [242]. Therefore, $ib = \sum_{i=1}^{n} s_i$, where n is the length of the list & s_i is the misinformation bias score of *ith* item in the list. Input bias is an unweighted bias, i.e it is not affected by the rank/ordering of the items. An *ib* of 1 indicates that all items in the list promote misinformation. By contrast, *ib* of -1 indicates that all items in the list debunk misinformation.

(ii) The *output bias* (*ob*) of a ranked list is the overall bias present in the SERPs and is the sum of biases introduced due to input and ranks of the input. It is computed as the cumulative weighted average of misinformation bias scores of items in the ranked list [242]. The score assigns more weight to the higher ranked items. I first calculate weighted bias score B(r) of every rank r, which is the average misinformation bias of results ranked from 1 to r. Thus, B(r) = $\frac{\sum_{i=1}^{r} s_i}{r}$, where s_i is the misinformation bias score of *ith* item. Output bias (ob) is the average of weighted bias score B(r) for all ranks. Thus, by definition ob = $\frac{\sum_{i=1}^{r} B(i)}{r}$.

(iii) The *ranking bias* (*rb*) is the bias introduced by the ranking algorithm of the search engine [242]. It is calculated by subtracting input bias from output bias. Thus, *rb* = *ob-ib*. In this case, high-ranking bias indicates that the search algorithm ranks misinformative products higher than neutral or debunking products.

Why do I need three bias scores? Amazon's search algorithm is not only selecting the products to be shown in the search results but it is also ranking them according to their internal algorithm. Therefore, the overall bias (ob) could be introduced either at the product selection stage (ib), or ranking stage (rb) or both. Studying all three biases gives us an elaborate understanding of how biases are introduced by the search algorithm. All three bias values (*ib*, *ob* and *rb*) lie between -1 and 1. A bias score larger than 0 indicates a lean towards misinformation. Conversely, a bias score less than 0

Rank r	Items	Bias of each product	Bias till rank r	Bias value
1	i_1	s_1	B(1)	s_1
2	i_2	s_2	B(2)	$\frac{1}{2}(s_1 + s_2)$
3	i_3	s_3	B(3)	$\frac{1}{3}(s_1 + s_2 + s_3)$
Input Bias (ib)				$\frac{1}{3}(s_1 + s_2 + s_3)$
Output Bias (ob)				$\frac{1}{3}\,[s_1(1 + \frac{1}{2} + \frac{1}{3}) + s_2(\frac{1}{2} + \frac{1}{3}) + s_3(\frac{1}{3})]$
Rank Bias (rb)				ob-ib

Table 5.6: Example illustrating the bias calculations. For a given query, Amazon's search engine presents users with the following products in the search results i_1, i_2 and i_3. The misinformation bias scores of the products are s_1, s_2 and s_3 respectively. The table has been adopted from previous work [242]. A bias score larger than 0 indicates a lean towards misinformation.

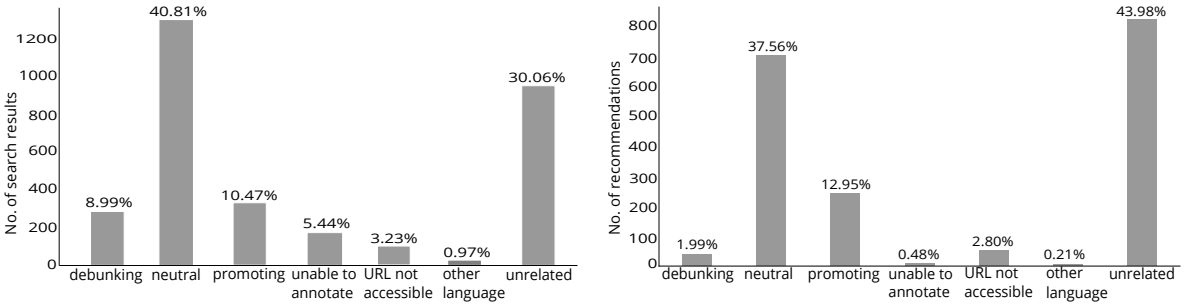
indicates a propensity towards debunking information. I only consider top 10 search results in each SERP. Thus, in the bias calculations, rank always varies from 1 to 10.

5.5 RQ1 Results [Unpersonalized audit]: Quantify misinformation bias

The aim of the *Unpersonalized audit* is to determine the amount of bias in search results. Below I present the input, rank, and output bias detected by the audit in search results of all 10 vaccine-related topics with respect to 5 search filters.

5.5.1 RQ1a: Search results

I collected 36,000 search results from the *Unpersonalized audit* run, out of which 3,180 were unique. Recall, I collected these products by searching for 48 search queries belonging to vaccine-related topics and sorting results by each of the 5 Amazon filters. I later extracted and annotated top 10 search results from all the collected SERPs resulting in 3,180 annotations. Figure 5.7a shows the number (and percentage) of products corresponding to each annotation value. Through the audits, I find a high percentage (10.47%) of misinformative products in the search results. Moreover, the number of misinformative products outnumbered the debunking products. Figure 5.8 illustrates the distribution of categories of Amazon products annotated as debunking (-1), neutral (0) and promoting (1). Note that the products promoting health misinformation primarily belong to categories Books (35.43%), Kindle eBooks (28.52%),



(a) Search results **(b)** Recommendations

Figure 5.7: RQ1a: (a) Number (percentage) of search results belonging to each annotation value. While majority of products have a neutral stance (40.81%), products promoting health misinformation (10.47%) are greater than products debunking health misinformation (8.99%). (b) Number (percentage) of recommendations belonging to each annotation value. A high percentage of product recommendations promote misinformation (12.95%) while percentage of recommendations debunking health misinformation is very low (1.99%).

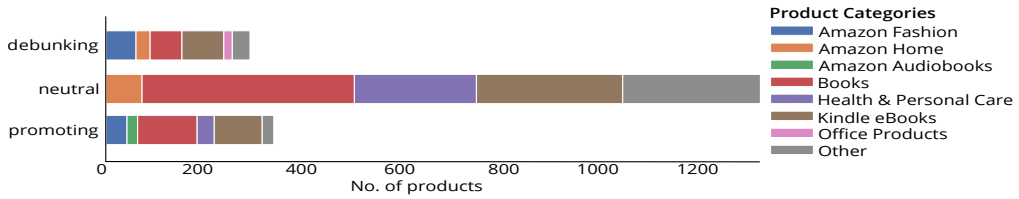


Figure 5.8: RQ1a: Figure showing categories of promoting, neutral and debunking Amazon products (search results). All categories occurring less than 5% were combined and are presented as *other* category. Note that misinformation exists in various forms on Amazon. Products promoting health misinformation include books (Books, Kindle eBooks, Audible Audiobooks), apparel (Amazon Fashion) and dietary supplements (Health & Personal Care). Additionally, proportion of books promoting health misinformation is much greater than proportion of books debunking misinformation.

[Categories of debunking, neutral and promoting Amazon products] Debunking products mostly belong to categories, Kindle eBooks, Books, Amazon fashion and Amazon home. Neutral products mostly belong to categories Books, Kindle eBooks, Health & Personal care and Amazon home. Promoting products belong to categories Books, Kindle eBooks, Health & Personal care and Amazon fashion.

Amazon Fashion (12.61%)—a category that includes t-shirts, apparel, etc. and Health & Personal Care (10.21%)—a category consisting of dietary supplements. Below I discuss the misinformation bias observed across all the vaccine-related topics, the Amazon search filters and search queries.

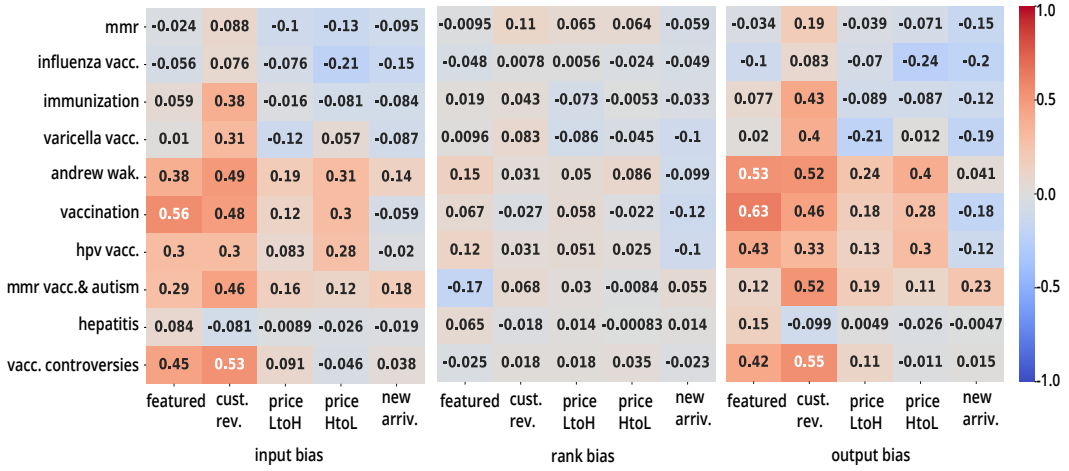


Figure 5.9: RQ1a: Input, rank and output bias for all 10 vaccine-related topics across five search filters. The bias scores are average of scores obtained for each of the 15 days. Input and rank bias is positive (>0) in the search results of majority of topics for filters "featured" and "average customer review". A bias value greater than 0 indicates a lean towards misinformation. Topics "andrew wakefield" and "mmr vaccine & autism" have a positive input bias across all five filters indicating that search results of these topics contain large number of products promoting health misinformation irrespective of the filter used to sort the search results. Topic "vaccination" has the highest overall bias (output bias) of 0.63 followed by topic "andrew wakefield" that has output bias of 0.53.

5.5.1.1 Misinformation bias in vaccine related topics

I calculate the input, rank and output bias for each of the 10 search topics. All the bias scores presented are average of scores obtained across the 15 days of audit. The bias score for a topic is also the average across each of the constituting search queries. Figure 5.9 shows the bias scores for all the topics, search filters and bias combinations.

Input bias: I observe a high input bias (>0) for all topics except "hepatitis" for the "average customer review" filter indicating presence of large number of misinformative books in the SERPs when search results are sorted by this filter. Similarly, input biases for most topics is also positive for "featured" filter. Note, "featured" is the default Amazon filter. Thus, by default Amazon is presenting more misinformative search results to users searching for vaccine related queries. Topics "andrew wakefield", "vaccination" and "vaccine controversies" have highest input biases for the both "featured" and "average customer review" filters. Another noteworthy trend is the negative input bias for 7 out of 10 topics with respect to filter "newest arrivals" indicating that there are more debunking products present in the SERP when users look for newly

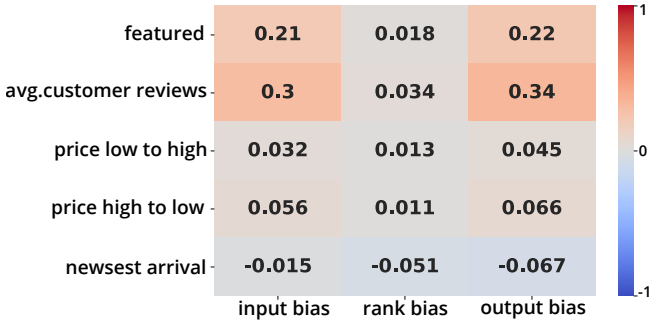


Figure 5.10: Input, rank and output bias for all filter types.

appearing products on Amazon. "andrew wakefield" and "mmr vaccine & autism" are the only two topics that have the high input bias (>0) across all the five filters. Interestingly, there is no topic that has negative input bias across all filters. Recall, a negative (<0) bias indicates a debunking lean. Topics "mmr" and "hepatitis" have negative bias scores in four out of five filters.

Rank bias: 8 out of 10 topics have positive rank bias for filters "price low to high" and "average customer reviews" and 6 out of 10 topics have positive rank bias for filter "featured". These results suggest that Amazon's ranking algorithm favors misinformative products and ranks them higher when customers filter their search results by the aforementioned filters. Some topics have negative input bias but positive rank bias. Consider topic "mmr" with respect to filter "price low to high" whose input bias is -0.1 but the rank bias is 0.065. This observation suggests that although the SERPs obtained had more debunking products, a few misinformative products were still ranked higher. Rank bias for 8 out of 10 topics with respect to filter "newest arrivals" was negative, similar to what I observed for input bias.

Output bias: Output bias is positive (>0) for most of the topics with respect to filters "featured" and "average customer reviews". Recall, a bias value greater than 0 indicates a lean towards misinformation. Topic "vaccination" has the highest output bias value of 0.63 for filter "featured". On the other hand, topic "hepatitis" has least output bias for filter "newest arrivals".

5.5.1.2 Misinformation bias in search filters

Figure 5.10 shows the results for all 5 filters. Bias scores are averaged across all search queries. All 5 filters except "newest arrivals" have positive input, rank, and output

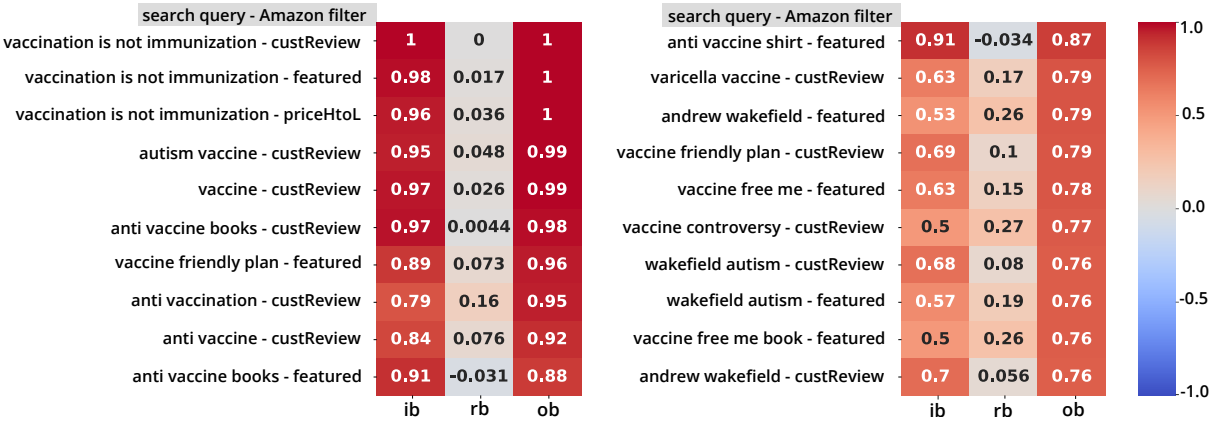
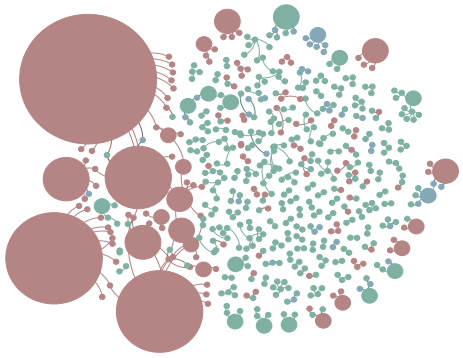


Figure 5.11: Top 20 search query-filter combinations with highest output bias. In other words, these query-filter combinations are the most problematic ones containing highest amount of misinformation.

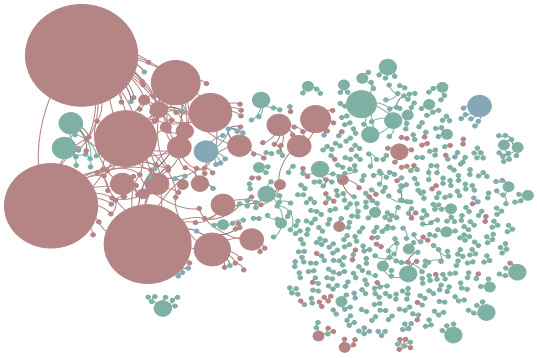
misinformation bias. Filter "average customer review" has the highest positive bias indicating that misinformative products belonging to vaccine related topics receive higher ratings. I present the implications of these results in the Discussion section.

5.5.1.3 Misinformation bias in search queries

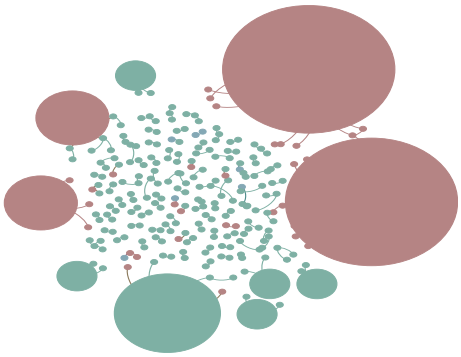
Figure 5.11 shows the top 20 search queries and filter combinations with highest output bias. Predictably, filter "newest arrivals" does not appear in any instance. Surprisingly, 9 search query-filter combinations have very high output biases (ob > 0.9). Search query "vaccination is not immunization" has output bias of 1 for three filter types. Most of the search queries in Figure 5.11 have a negative connotation, i.e the queries themselves have a bias (e.g search queries anti vaccine books, vaccination is not immunization indicates an intent to search for misinformation). This observation indicates that if you search for anti vaccine stuff, you will get high amount of vaccine and health misinformation. The most troublesome observation is the presence of high output bias for generic and neutral search queries, "vaccine" (ob = 0.99) and "varicella vaccine" (ob = 0.79). These results indicate that, unlike companies like Pinterest, who have altered their search engines in response to vaccine related queries [383], Amazon has not made any modification to its search algorithm to push less anti vaccine products to users.



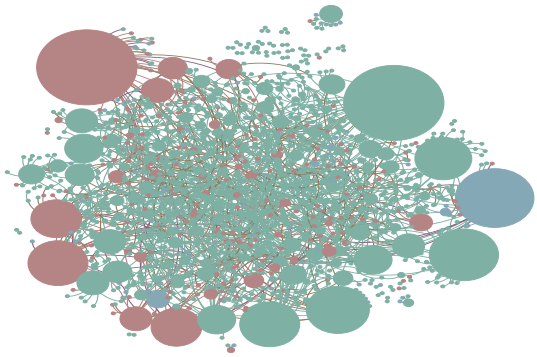
(a) Customers who bought this item also bought (CBB)



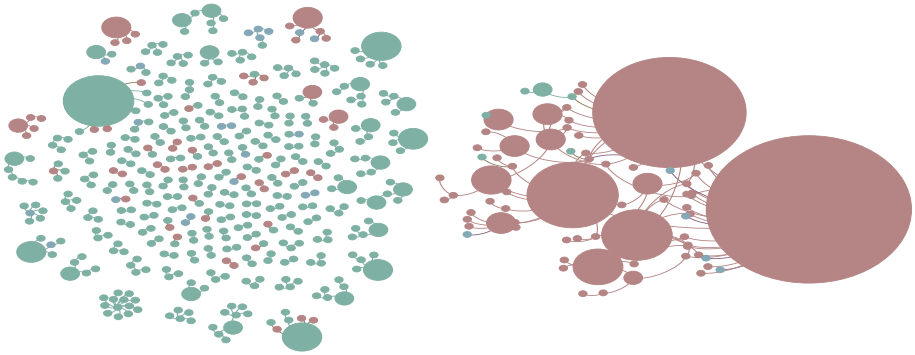
(b) Customers who viewed this item also viewed (CVV)



(c) Frequently bought together (FBT)



(d) Sponsored products related to this item



(e) What other items customers buy after viewing this item (CBV). Note that the recommendation graph for CBV recommendation type is indeed one figure. It consists of two disconnected components, indicating strong filter bubble effect.

Figure 5.12: Recommendation graphs for 5 different types of recommendations collected from the product pages of top three search-results obtained in response to 48 search queries, sorted by 5 filters over a duration of 15 days during *Unpersonalized audit* run. ■ denotes products annotated as misinformative, ■ as neutral and ■ as debunking. Node size is proportional to the times the product was recommended in that recommendation type. Large sized red nodes coupled with several interconnections between red nodes indicate a strong filter-bubble effect where recommendations of misinformative products returned more misinformation.

5.5.2 RQ1b: Product page recommendations

I extracted the product page recommendations of top 3 search results present in the SERPs. The product page constitutes various types of recommendations. For analysis, I consider the first product present in 5 types of recommendations "Customers who bought this item also bought", "Customers who viewed this item also viewed", "Frequently bought together", "Sponsored products related to this item" and "What other items customers buy after viewing this item". The process resulted in 16,815 recommendations out of which 1,853 were unique. Figure 5.7b shows the number and percentage of recommendations belonging to different annotation values. The percentage of misinformative recommendations (12.95%) is much higher than the debunking recommendations (1.95%). The total input bias in all 16,815 recommendations is 0.417 while in all 1,853 unique recommendations is 0.109, indicating a lean towards misinformation.

Does filter-bubble effect occur in product page recommendations? To answer, I compared the misinformation bias scores of all types of recommendations considered together (refer Table 5.7). Kruskal Wallis Anova test revealed the difference to be significant (KW H(2, N=16815) = 6,927.6, p=0.0). Post-hoc Tukey HSD test showed that the product page recommendations of misinformative products contain more misinformation when compared to recommendations of neutral and debunking products. Even more concerning is that the recommendations of debunking products have more misinformation than neutral products. To investigate further I qualitatively studied the recommendation graphs of each of the 5 recommendation types (Figure 5.12). Each node in the graph represents an Amazon product. An edge A→ B indicates that B was recommended in the product page of A. Node size is proportional to the number of times the product was recommended.

5.5.2.1 Recommendation type- Customers who bought this item also bought (CBB)

Misinformation bias scores of CBB recommendations are significantly different for debunking, neutral, and promoting products (KW H(2, N=3133) = 2136.03, p=0.0). Post hoc tests reveal that CBB recommendations of misinformative products have more misinformation when compared to CBB recommendations of neutral and debunking products. Additionally CBB recommendations of neutral products have more misinformation than CBB recommendations of debunking products. The findings are evident from Figure 5.12a too. For example, there are several instances of red nodes connected

Type of product page recommendations	Kruskal Wallis Anova Test	Post hoc Tukey HSD	d	n	m
All	KW H(2, N=16815) = 6,927.6, p=0.0	M>D & M>N & D>N	37	1576	240
Cust. who bought this item also bought (CBB)	KW H(2, N=3133) = 2136.03, p=0.0	M >D & M>N & N>D	11	225	66
Cust. who viewed this item also viewed (CVV)	KW H(2, N=6575) = 628.52, p=3.2e-137	M>D & M>N & D>N	18	331	100
Frequently bought together (FBT)	KW H(2, N=2234) = 1611.34, p=0.0	M>D & M>N & D>N	1	111	16
Sponsored products related to this item	KW H(2, N=388) = 277.08, p=6.8e-61	M>D & M>N	7	953	98
What other items cust. buy after viewing this item (CBV)	KW H(2, N=4485) = 2673.95, p=0.0	M>D & M>N & D>N	9	230	57

Table 5.7: RQ1b: Analyzing echo chamber effect in product page recommendations. M, N and D are the means of misinformation bias scores of products recommended in the product pages of misinformative, neutral and debunking Amazon products respectively. Higher means indicate that recommendations contain more misinformative products. For example, M>D indicates that recommendations of misinformative products have more misinformation than recommendations of debunking products. d, n and m are number of unique products annotated as debunking, neutral and promoting for each recommendation type.

to each other. In other words, if you click on a misinformative search result, you will get misinformative products in CBB recommendations. Few of the green nodes are attached to red ones indicating that CBB recommendation of a neutral product sometimes contain a misinformative product. The most recommended product present in CBB is a misinformative Kindle book titled *Miller's Review of Critical Vaccine Studies: 400 Important Scientific Papers Summarized for Parents and Researchers* (B07NQW27VD).

5.5.2.2 Recommendation type- Customers who viewed this item also viewed (CVV)

Misinformation bias scores of CVV recommendations are significantly different for debunking, neutral and promoting products (KW H(2, N=4485) =2673.95, p=0.0) . Post hoc test indicates that CVV recommendations of misinformative products have more misinformation than CVV recommendations of debunking and neutral products. Notably, CVV recommendations of debunking products contain more misinformation than CVV recommendations of neutral products. In the recommendation graph (Figure 5.12b), I see edges connecting multiple red nodes supporting my finding that CVV recommendations of misinformative products mostly contain other misinformative products. The most recommended product in this recommendation type is a misinformative Kindle book titled *Dissolving Illusions* (B00E7FOA0U).

5.5.2.3 Recommendation type- Frequently bought together (FBT)

Misinformation bias scores of FBT recommendations are significantly different for debunking, neutral and promoting products (KW H(2, N=2234) = 1611.34, p=0.0). Post hoc tests reveal that amount of misinformation in FBB recommendations of misinformative products is significantly more than the FBB recommendations of neutral and debunking products. The finding is also evident from the graph (Figure 5.12c). There are large sized red nodes attached to other red nodes and several green nodes attached together indicating the presence of a strong filter-bubble effect. "Frequently bought together" can be considered an indicator of buying patterns on the platform. The post hoc tests indicate that people buy multiple misinformative products together. The most recommended product in this recommendation type is a misinformative Paperback book titled *Dissolving Illusions: Disease, Vaccines, and The Forgotten History* (1480216895).

5.5.2.4 Recommendation type- Sponsored products related to this item

Most of the sponsored recommendations are either neutral or promoting (Figure 5.12d). Statistical test reveals that the misinformation bias score of sponsored recommendations are significantly different among debunking, neutral and promoting products (KW H(2, N=6575) = 628.52, p=3.2e-137). Post hoc tests reveal same results as for CVV recommendations. There are two most recommended sponsored books. First is a misinformative paperback book titled *Vaccine Epidemic: How Corporate Greed, Biased Science, and Coercive Government Threaten Our Human Rights, Our Health, and Our Children* (1620872129). Second is a neutral Kindle book titled *SPANISH FLU 1918: Data and Reflections on the Consequences of the Deadliest Plague, What History Teaches, How Not to Repeat the Same Mistakes* (B08774MCVP).

5.5.2.5 Recommendation type- What other items customers buy after viewing this item (CBV)

Misinformation bias scores of CBV recommendations are significantly different for debunking, neutral and promoting products (KW H(2, N=2234) = 1611.34, p=0.0). Post hoc tests reveal a filter-bubble effect in the product recommendations. CBV recommendations of misinformative products contain more misinformation than neutral or debunking products. Furthermore, CBV recommendations of debunking products contain more misinformation than neutral products. This is troubling since

Actions performed to build account history	RQ2a												RQ2b									RQ2c		
	Search results												Recommendations									Auto complete suggestions		
	Featured			Avg. customer reviews			Price low to High			Newest Arrivals			Homepage			Pre-purchase			Product page					
	D	N	M	D	N	M	D	N	M	D	N	M	D	N	M	D	N	M	D	N	M	D	N	M
Search product	IR	IR	IR	NP	NP	NP	NP	NP	NP	NP	NP	NP	-	-	-	X	X	X	X	X	X	NP	NP	NP
Search & click product	IR	IR	IR	NP	NP	NP	NP	NP	NP	NP	NP	NP	KW H(2, N=42) = 32.07, p = 1.08e-07 M>N>D			X	X	X	KW H(2, N=42) = 24.89, p = 3.94e-06 M>D & M>N			NP	NP	NP
Search + click & add to cart product	IR	IR	IR	NP	NP	NP	NP	NP	NP	NP	NP	NP	KW H(2, N=42) = 33.48, p = 5.38e-08 M>N>D			KW H(2, 42) = 32.63, p = 8.19e-08 M>N>D			KW H(2, N=42) = 24.05, p = 5.98e-06 M>D & M>N			NP	NP	NP
Search + click & mark "Top rated, All positive review" as helpful	IR	IR	IR	NP	NP	NP	NP	NP	NP	NP	NP	NP	KW H(2, N=42) = 32.33, p = 9.52e-08 M>N>D			X	X	X	KW H(2, 42) = 23.36, p = 8.44e-06 M>N & M>D			NP	NP	NP
Following contributor	IR	IR	IR	NP	NP	NP	NP	NP	NP	NP	NP	NP	-	-	-	X	X	X	X	X	X	NP	NP	NP
Search product on Google	IR	IR	IR	NP	NP	NP	NP	NP	NP	NP	NP	NP	-	-	-	X	X	X	X	X	X	NP	NP	NP

Table 5.8: RQ2: Table summarizing RQ2 results. IR suggests noise and inconclusive results, i.e search results of control and its twin seldom matched. Thus, difference between treatment and control could either be attributed to noise or personalization, making it impossible to study the impact of personalization on misinformation. NP denotes little to no personalization. - indicates that the given activity had no impact on the component. X indicates that component was not collected for the activity. M, N and D indicate average per day bias in the component collected by accounts that built their history by performing actions on misinformative, neutral or debunking products. Higher mean value indicates more misinformation. For example, consider the cell corresponding to action "search + click & add to cart product" and "Homepage" recommendation. M>N>D indicates that accounts adding misinformative products to cart ends up with more misinformation in their homepage recommendations in comparison to accounts that add neutral or debunking products to cart.

users who are clicking on products that present scientific information are pushed more misinformation in this recommendation type.

The presence of an echo chamber is quite evident in the recommendation graph (see Figure 5.12e). The graph has two disconnected components, one comprising a mesh of misinformative products indicating a cluster of misinformative products that keep getting recommended. CBV is also indicative of buying patterns of Amazon users. The algorithm has learnt that people viewing misinformative products end up purchasing them. Thus, it pushes more misinformative items to users that click on them, creating a feedback loop. The most recommended product in this recommendation type is a misinformative Kindle book titled *Miller's Review of Critical Vaccine Studies: 400 Important Scientific Papers Summarized for Parents and Researchers* (B07NQW27VD).

5.6 RQ2 Results [Personalized audit]: Effect of personalization

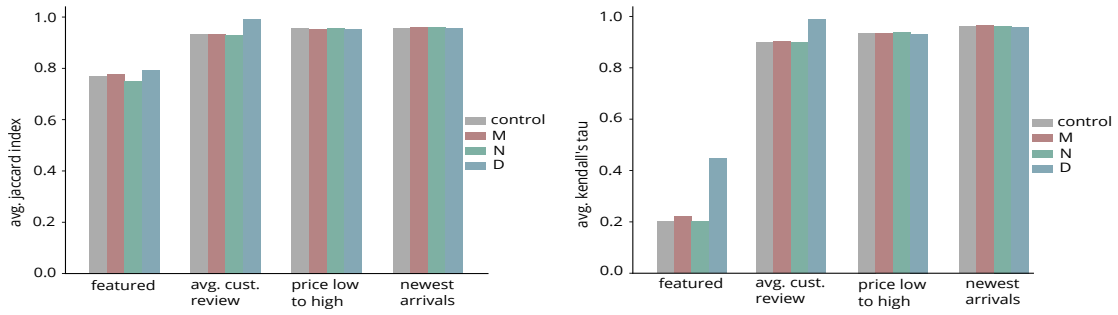
The aim of the Personalized audit was to determine the effect of personalization due to account history on the amount of misinformation returned in search results and various recommendations. Table 5.8 provides a summary. Below, I explain the effect of personalization on each component.

5.6.1 RQ2a: Search Results

I measure personalization in search results for each Amazon filter using two metrics: Jaccard index and Kendall τ coefficient. Jaccard index determines similarity between two lists. A Jaccard index of 1 indicates that the two lists have same elements and zero indicates that the lists are completely different. On the other hand, Kendall τ coefficient, also known as Kendall rank correlation coefficient determines the ordinal correlation between two lists. It can take values between [-1,1] with -1 indicating that lists have inverse ordering, 0 signifying no correlation and 1 suggesting that items in the list have same ranks.

First I compare search results of control account and its twin. Recall I created twins for the two control accounts in the *Personalized audit* to establish the baseline noise. Ideally, both should have Jaccard and Kendall rank correlation coefficient closer to 1 since the accounts do not build any history, are set up in a similar manner, perform searches at the same time and are in the same geolocation. Next, I compare search results of control account with treatment accounts that built account histories by performing different actions. If personalization is occurring, the difference between search results of treatment and control should be more than the baseline noise (or Jaccard index and Kendall τ should be less). Whereas, if the baseline noise itself is large, it indicates inconsistencies and randomness in the search results. Interestingly, I found significant noise in search results of control and its twin for "featured" filter with jaccard index <0.8 and Kendall's rank correlation coefficient <0.2, that is, control and its twins seldom matched. Presence of noise suggests that Amazon is injecting some randomness in the "featured" search results. Unfortunately, this means that I would be not be able to study the effect of personalization on the accounts for the "featured" search filter setting.

For the other three search filters, "average customer review", "price low to high"



(a) **(b)**

Figure 5.13: Investigating the presence and amount of personalization due to "following contributors" action by calculating (a) Jaccard index and (b) kendall's tao metric between search results of treatment and control. M, N and D indicate results for accounts that follow contributors of misinformative, neutral and debunking products respectively.

and "newest arrivals", I see high (>0.8) jaccard index and kendall τ metric values between and control and its twin. Additionally, I do not see any personalization for these filters since metrics values for treatment-control comparison are similar to that of control-twin comparison. Figure 5.13 shows the metrics calculation for control account and treatments that have built their search histories by following contributor's of misinformative, neutral and debunking products. I see two minor inconsistencies for filter "average customer review" in accounts building their history on debunking products. The metric values for treatment-control account was higher than control-twin value. This means treatment received more similar results to control than its twin account. In any case, the treatment account does not see more inconsistency than the control and its twin indicating no personalization. Other user actions show similar results, hence, I have removed their results for brevity.

5.6.2 RQ2b: Recommendations

I investigated the occurrence of personalization and its impact on the amount of misinformation in three different types of recommendations. I discuss each type of recommendation below.

Homepage recommendations: I find that homepages are personalized only when a user performs click actions on the search results. Thus, actions "add to cart", "search + click" and "mark top rated most positive review helpful" led to homepage personalization. On the other hand, homepages were not personalized for actions "follow

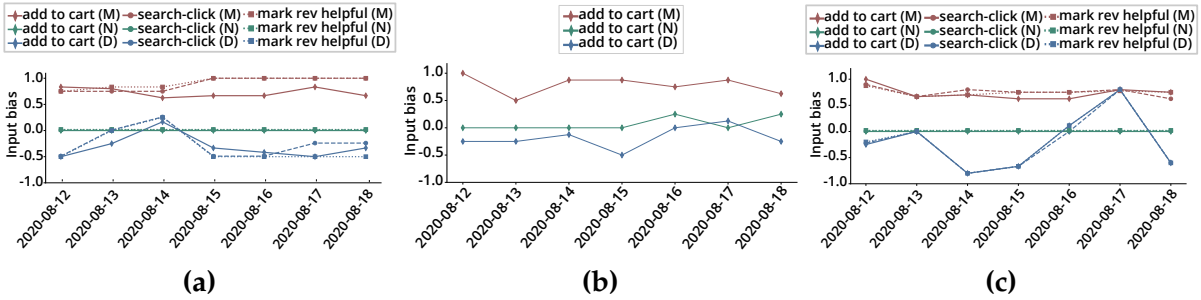


Figure 5.14: (a) Input bias in homepages of accounts performing actions 'add to cart", "search + click" and "mark top rated all positive review" for seven days of experiment run. (b) Input bias in pre-purchase recommendations of accounts for 7 days experiment run. These recommendations are only collected for accounts adding products to their carts. (c) Input bias in product pages of accounts performing actions "add to cart", "search + click" and "mark top rated all positive review" for 7 days of experiment run. M, N and D indicate that the accounts performed actions on misinformative, neutral and debunking products respectively.

contributor", "search product" and "google search" actions. After identifying the actions leading to personalized homepages, I investigate the impact of personalization on the amount of misinformation. In other words, I investigate how misinformation bias in homepages is different for accounts building their history by performing actions on misinformative, neutral and debunking products. For each action, I had 6 accounts, two replicates for each action and product type (misinformation, neutral and debunking). For example, for action "add to cart" two accounts built their history by adding misinformative products to cart for 7 days, two added neutral products and two accounts added debunking products to their carts. I calculate per day input bias (ib) in homepages by averaging the misinformation bias scores of each recommended product present in the homepage. Therefore, for every account I have seven bias values. I consider only top two products in each recommendation type. Recall, homepages could contain three different types of recommendations 'Inspired by your shopping trends", "Recommended items other customers often buy again" and "Related to items you've viewed". All the different types are considered together for analysis.

Statistical tests reveal significant differences in the amount of misinformation present in homepages of accounts that built their histories by performing actions on misinformative, neutral and debunking products (see Table 5.8). This observation holds true for all three activities "add to cart", "search + click" and "mark top rated most positive review helpful". Post hoc test reveal an echo chamber effect. Amount of misinformation in recommendations of products performing actions on misinforma-

tive products is more than the amount of misinformation in homepages of accounts performing actions on neutral products which in turn is more than the misinformation present in homepages of accounts performing actions on debunking products.

Figure 5.14a shows per day input bias of homepages of different accounts performing different actions. I take an average of the replicates for plotting the graph. Surprisingly, performing actions "mark top rated most positive review helpful" and "search + click" on a misinformative product leads to highest amount of misinformation in the homepages, even more than the homepages of accounts adding misinformative products to the cart. This means that amount of misinformation present in homepage is comparatively less once a user shows an intention to purchase a misinformative product but high if a user shows interest in the misinformative product but doesn't show an intention to buy it. Figure 5.14a also shows that amount of misinformation present in homepages of accounts performing actions "mark top rated most positive review helpful" and "search + click" on misinformative products gradually increases and becomes 1 on day 4 (2020-08-15). Bias value 1 indicates that all analysed products in homepages were misinformative. Homepage recommendations of products performing actions on neutral objects show 0 bias constantly indicating all recommendations on all days were neutral. On the other hand, average bias in homepages of accounts building history on debunking accounts rises a little above 0 in the first three days but eventually fells below 0 indicating a debunking lean.

Pre-purchase recommendations: These recommendations are only presented to users that add product(s) to their Amazon cart. Therefore, they were collected for 6 accounts, 2 of which added misinformative products to cart, 2 added neutral products and the other 2 added debunking products. These recommendations could be of several types. See Figure 5.1b for an example of pre-purchase page. For the analysis, I consider the first product present in each recommendation type. Statistical tests reveal significant difference in the amount of misinformation present in pre-purchase recommendations of accounts that added misinformative, neutral and debunking products to cart (KW $H(2, 42) = 32.63, p = 8.19e\text{-}08$). Those adding misinformative products to cart contain more misinformation than the accounts adding neutral or debunking products to their carts. Figure 5.14b shows the input bias in the pre-purchase recommendations for all the accounts. There is no coherent temporal trend, indicating that the input bias in this recommendation type depends on the particular product being added to cart. However, an echo chamber effect is evident. For example, bias in pre-purchase

recommendations of accounts adding misinformative products to cart is above 0 for all 7 days.

Product recommendations: I collect product recommendations for accounts performing actions "add to cart", "search + click" and "mark top rated most positive review helpful". I find significant difference in the amount of misinformation present in product page recommendations when accounts performing the aforementioned actions on misinformative, neutral and debunking products (refer Table 5.8). Post hoc analysis reveals that product page recommendations of misinformative products contain more misinformation than those of neutral and debunking products. Figure 5.14c shows the input bias present in product pages for various accounts. The bias for neutral products is constantly 0 across the 7 days, but for misinformative products, it is constantly greater than 0 for all actions. I see an unusually high bias value on the 6th day (2020-08-17) of the experiment for accounts performing actions on debunking product titled *Reasons to Vaccinate: Proof That Vaccines Save Lives* (B086B8MM71). I checked the product page recommendations of this particular debunking book and found several misinformative recommendations on its product page.

5.6.3 RQ2c: Auto-complete suggestions

I audited auto-complete suggestions to investigate how personalization affects the change in search query suggestions. My initial hypothesis was that performing actions on misinformative products could increase the auto-complete suggestions of anti-vaccine search queries. However, I found little to no personalization in the auto-complete suggestions indicating that account history built by performing actions on vaccine-related misinformative, neutral, or debunking products have little to no effect on how auto-complete suggestions of accounts change. In the interest of brevity, I do not add the results and graphs for this component.

5.7 Discussion

There is a growing concern that e-commerce platforms are becoming hubs of dangerous medical misinformation. Because of a lack of regulatory policies, websites like Amazon are providing a platform to people who are making money by selling misinformation—dangerous anti-vaccine ideas, pseudoscience treatments, or un-

proven dietary alternatives—some of which could have dangerous effects on people's health and well-being. With a US market share of 49%, Amazon is the leading product search engine in the United States [112]. Thus, any misinformation present in its search and recommendations could have a far-reaching influence where they can negatively shape users' viewing and purchasing patterns. Therefore, in this study, I audited Amazon for the most dangerous form of health misinformation—vaccine misinformation. My work resulted in several critical findings with far-reaching implications. I discuss them below.

5.7.1 Amazon: a marketplace of multifaceted health misinformation

The analysis shows that Amazon hosts a variety of health misinformative products. Maximum number of such products belong to the category Books and Kindle eBooks (Figure 5.8). Despite the enormous amount of information available online, people still turn to books to gain information. A Pew Research survey revealed that 73% of Americans read atleast one book in a year [309]. Books are considered "intellectual heft", have more presence than scientific journals and thus, leave "a wider long lasting wake" [197]. Thus, anti-vaccine books could have a wider reach and can easily influence the audience negatively. Moreover, it does not help that a large number of anti-vaccine books are written by authors with medical degrees [358]. Not just anti-vaccine books, there are abundant pseudoscience books on the platform, all suggesting unproven methods to cure diseases. I found diet books suggesting recipes with colloidal silver—an unsafe product, as an ingredient. Some of the books proposing cures for incurable diseases, like autism and auto immune diseases, can have a huge appeal for people suffering with such diseases [328]. Thus, there is an urgent need to check the quality of health books presented to the users.

The next most prominent category of health misinformative products is Amazon Fashion. Numerous apparel are sold on the platform with innovative anti-vaccine slogans, giving tools to the anti-vaccine propagandists to advocate their anti-vaccine agenda and gain visibility, not just in the online world, but in the offline world. During the annotation process, I also found many dietary supplements claiming to treat and cure diseases—a direct violation of Amazon's policy on dietary supplements. Overall, I find that health misinformation exists on the platform in various forms—books, t-shirts, and other merchandise. Additionally, it is very easy to sell problematic content because of the lack of appropriate quality-control policies and their enforcement.

5.7.2 Amazon search results: a stockpile of health misinformation

Analysis of the *Unpersonalized audit* revealed that 10.47% of search results promote vaccine and other health-related misinformation. Notably, the higher percentage of products promoting misinformation compared to debunking suggests that anti-vaccine and problematic health-related content is churned out more and the attempts to debunk the existing misinformation is less. I also found that Amazon's search algorithm puts more health misinformative products in search results than debunking products leading to high input bias for topics like "vaccination", "vaccine controversies", "hpv vaccine", etc. This is specifically true for search filters "featured" and "average customer reviews". Note, that "featured" is the default search filter indicating that by default users will see more misinformation when they search for the aforementioned topics. On the other hand, if users want to make a purchase decision based on product ratings, again users will be presented with more misinformation because it seems misinformative products have higher user ratings on the platform. I also found a ranking bias in Amazon's search algorithm with misinformative products ranked higher. Past research has shown that people trust higher ranked search results [181]. Thus, more number of higher ranked misinformative products can make problematic ideas in these products appear mainstream. The only positive finding of my analysis was the presence of more debunking products in search results sorted by filter "newest arrivals". This might indicate that higher quality products are being sold on the platform in recent times. However, since there are no studies/surveys indicating which search filters are mostly used by people while making purchase decisions, it is difficult to conclude how beneficial this finding is.

5.7.3 Amazon recommendations: problematic echo chambers

Many search engines and social media platforms employ personalization to enhance users' experience on their platform by recommending them items that the algorithm think they will like based on their past browsing or purchasing history. But on the downside, if not checked, personalization can also lead users into a rabbit hole of problematic content. My analysis of *Personalized audit* revealed that an echo chamber exists on Amazon where users performing real-world actions on misinformative books are presented with more misinformation in various recommendations. Just a single click on an anti-vaccine book could fill your homepage with several other similar anti vaccine books. And if you proceed to add that book in your cart, Amazon again presents

more anti-vaccine books, nudging you to purchase even more problematic content. The worst discovery is that your homepages get filled with more misinformation if you just show an interest in a misinformative product (by clicking on it) compared to when you show an intention to buy it by adding product to your cart. Additionally on the product page itself, you are presented 5 different kinds of recommendations each of them presenting you with equally problematic content. In a nutshell, once you start engaging with misinformative products on the platform, you will be presented with more misinformative stuff at every point of your Amazon navigation route and at multiple places. These findings would not have been concerning if buying a milk chocolate would lead to recommendations of other chocolates of different brands. The problem is that Amazon is blindly applying its algorithms on all products including problematic content. Its algorithms do not differentiate or gives special significance to vaccine-related topics. Amazon has learnt from users' past viewing and purchasing behaviour and has categorized all the anti-vaccine and other problematic health cures together. It presents the problematic content to users performing actions on any of these products, creating a dangerous recommendation loop in the process. There is an urgent need for the platform to treat vaccine and other health related topics differently and ensure high quality searches and recommendations. In the next section, I present a few ways, based on my findings, that could assist the platform in combating health misinformation.

5.7.4 Combating health misinformation

Tackling online health misinformation is a complex problem and there is no easy silver-bullet solution to curb its spread. However, the first step towards addressing is accepting that there is a problem. Many Tech giants have acknowledged their social responsibility in ensuring high quality in health-related content and are actively taking many steps to ensure the same. For example, Google's policy "Your Money Or Your Life" classifies medical and health-related search pages as pages of particular importance, whose content should come from reputable websites [270]. Pinterest completely hobbled the search results some queries such as 'anti-vax' [383] and limited the search results for other vaccine-related queries to content from officially recognized health institutions [206]. Even Facebook—a platform known to have questionable content moderation policies—banned anti-vaccine ads and demoted the anti-vaccine content in its search results to make its access difficult [268]. Therefore, given the massive reach and user base of Amazon—206 million website visits every month

[34]—it is disconcerting to see that Amazon has not yet joined the bandwagon. To date, it has not taken any concrete steps toward addressing the problem of anti-vaccine content on its platform. I recommend several short-term and long-term strategies that the platform can adopt.

5.7.4.1 Short-term strategies: design interventions,

The simplest short-term solution would be to introduce design interventions. The *Unpersonalized audit* revealed high misinformation bias in search results. The platform can use interventions as an opportunity to communicate to users the quality of data presented to them by signalling misinformation bias. The platform could introduce a bias meter or scale that signals the amount of misinformation present in search results every time it detects a vaccine-related query in its search bar. The bias indicators could be coupled with informational interventions like showing Wikipedia and encyclopedia links, that have already been proven to be effective in reducing traffic to anti-vaccine content [234]. The second intervention strategy could be to recognise and signal source bias. During the massive annotation process, I realized that several health misinformative books have been written by known anti-vaxxers like Andrew Wakefield, Jenny Mccarthy, Robert S. Mendelsohn, etc. I also present a list of authors who have contributed to most misinformative books in Table 5.3. Imagine a design where users are presented with a message "The author is a known anti-vaxxer and is known to write books that might contain health minformation" every time they click a book written by these authors. An another extreme short term solution could be to either enforce a platform-wide ban prohibiting sale of any anti-vaccine product or hobble search results for anti-vaccine search queries.

5.7.4.2 Long term strategies: algorithmic modifications and policy changes.

Long term interventions would include modification of search, ranking and recommendation algorithm. My investigations revealed that Amazon's algorithm has learnt problematic patterns through consumer's past viewing and buying patterns. It has categorized all products of similar stance together (see several edges connecting red nodes—products promoting misinformation in Figure 5.12). In some cases, it has also associated some misinformative products with neutral and debunking products (refer Figure 5.12) Amazon needs to "unlearn" this categorization. Additionally, the platform should incorporate misinformation bias in their search and recommendation algorithms to reduce the exposure to misinformative content. There is also an urgent need

to introduce some policy changes. First and foremost, Amazon should stop promoting health misinformative books by sponsoring them. I found 98 misinformative products in the sponsored recommendations indicating that today, anti-vaccine outlets can easily promote their products by spending some money. Amazon should also introduce some minimum quality requirements that should be met before a product is allowed to be sponsored or sold on its platform. It can employ search quality raters to rate the quality of search results for various health-related search queries. Google has already set an example with its extensive Search Quality Rating process and guidelines [7, 169]. In recent times Amazon introduced several policy and algorithmic changes including roll out of a new feature "verified purchase" to curb fake reviews problem on its platform [334]. Similar efforts are required to ensure product quality as well. Amazon can introduce a similar "verified quality" or "verified claims" tag with health-related products once they are evaluated by experts. Having a product base of millions of products can make any kind of review process tedious and challenging. Amazon can start by targeting specific health and vaccine related topics that are most likely to be searched. My work itself presents a list of most popular vaccine-related topics that can be used as a starting point. Finally, I hope my work acts as a call to action for Amazon and also inspires other vaccine and health audits on other platforms.

5.8 Limitations

This study is not without limitations. First, I only considered top products in each recommendation-type present on a page while determining bias of the entire page. Annotating and determining bias of all the recommendations occurring in a page would give a much more accurate logic of recommendation algorithms. However, past studies have shown that the top results receive the highest number of clicks, thus, are more likely to receive attention from users [114]. Second, search queries themselves have inherent bias. For example query 'anti vaccine t-shirt' suggests that user is looking for anti-vax products. Higher bias in search results of neutral queries is much worse than that of biased queries. I did not segregate the analysis based on search query bias. Although, I did notice two neutral search queries namely 'vaccine' and 'varicella vaccine' appearing in the list of most problematic search-query and filter combinations. Third, while I audited various recommendations present on the platform, I did not analyse the email recommendations—product recommendations present outside the platform. A journalistic report pointed that email recommendations could be contaminated

too if a user shows an interest in a misinformative product but leaves the platform without buying it [121]. I leave investigation of these recommendations to future work. Fourth, in the *Personalized audit*, accounts only built history for a week. Moreover, experiments were only run on Amazon.com. I plan to continue to run the experiments and explore features such as geolocation for future audits. Fifth, the audit study only targeted results returned in response to vaccine-related queries. Since, Amazon is a vast platform that hosts variety of products and sellers, I cannot claim that my results are generalizable for other misinformative topics or conspiracy theories. However, my methodology is generic enough to be applied to other misinformative topics. Lastly, another major limitation of the study is that in the *Personalized audits* account histories were built in a very conservative setting. Accounts performed actions on only one product each day. Additionally, the actions were only performed on products with the same stance. In real-world it will be tough to find users who only add misinformative products to their carts for seven days continuously. But in spite of this limitation, my study still provides a peek into the workings of Amazon's algorithm and has paved the way for future audits that could use my audit methodology and extensive qualitative coding scheme to perform experiments considering complex real-world settings.

5.9 Conclusion

In this study, I conducted two sets of audit experiments on a popular e-commerce platform, Amazon to empirically determine the amount misinformation returned by its search and recommendation algorithm. I also investigated whether personalization due to user history plays any role in amplifying misinformation. My audits resulted in a dataset of 4,997 Amazon products annotated for health misinformation. I found that search results returned for many vaccine-related queries contain a large number of misinformative products leading to high misinformation bias. Moreover, misinformative products are also ranked higher than debunking products. My study also suggests the presence of a filter-bubble effect in recommendations, where users performing actions on misinformative products are presented with more misinformation on their homepages, product page recommendations, and pre-purchase recommendations. I believe, my proposed methodology to audit vaccine misinformation can be applied to other platforms to investigate health-misinformation bias. Overall, my study brings attention to the need for search engines to ensure high standards and quality of results for health-related queries.

CHAPTER 6

IDENTIFYING WAYS TO SUPPORT FACT-CHECKING ONLINE MISINFORMATION

6.1 Introduction

Chapters 3, 4, and 5, focused on auditing online platforms for algorithmically curated misinformation. This chapter focuses on determining ways to support online fact-checking as a way to combat online misinformation. While a lot of research has been done to design scalable technological systems for fact-checking, such systems fail to have an impact on fact-checking in the real world [171] for primarily for two reasons. First, their design is treated as a technical solution to what is often seen as a purely technological problem. But fact-checking is a complex socially-situated technical phenomenon involving collaboration among multiple stakeholder groups at various stages of the process. Yet, current automated fact-checking systems rarely take into account the insights and needs of "the human"—stakeholder groups who are central to this process. Second, automated fact-checking systems are limited in their applicability. For instance, most systems are either restricted to verifying claims about very specific public statistics by matching them against official figures (e.g., unemployment rate, inflation rate, etc.) or they are limited to identifying simple declarative claims to debunk [171]. Hence, automated fact-checking solutions fail to generalize to real-world fact-checking scenarios [171]. In other words, the rigidity of a purely technical system lacks the social flexibility necessary to support an inherent

socio-technical process.

In the last decade, HCI and CSCW communities have developed a better understanding of the gap between the social and the technical [39, 129, 163, 395] and thus, are well positioned to develop an understanding of the socio-technical mechanisms underlying fact-checking. Yet, to date, we know very little about how fact-checking is done in practice and what we could do to socially and technically support the fact-checking process. In this work, I elucidate how fact-checking is practiced by laying bare the human and technological infrastructures that facilitate and shape the fact-checking process in a fact-checking team/organization. I attempt to foreground the social by revealing the synergistic collaboration that occurs among human infrastructure —various stakeholder groups that work together to accomplish fact-checking work. I provide visibility to the stakeholder groups' roles, needs, and activities, many of which often remain invisible to the external world. I also highlight the technological infrastructure —the tools, technology, processes, and policies—that supports and enables the work of the stakeholder groups. The foregrounding of the infrastructures supporting the fact-checking work helps us unravel the technical, policy, and information barriers to fact-checking. My hope is that by considering both the human and technological infrastructures underlying the fact-checking process, we might narrow the "design-reality gap" [195]—the gap that exists between the needs of stakeholder groups involved in fact-checking and design of technical systems for fact-checking. Overall, in this work, I answer the following research questions.

RQ1: What are the various infrastructures supporting fact-checking work?

> **RQ1a Human infrastructure:** Who are the various stakeholder groups involved in the fact-checking process? What roles do they play? How do they evaluate priorities and collaborate together to make decisions?
>
> **RQ2b Technological infrastructure:** How do tools, technology and policy support stakeholder groups in performing their roles?

RQ2 Barriers to fact-checking: What are the various needs and challenges of stakeholder groups involved in the fact-checking process?

To answer the research questions, I adopt a multi-stakeholder approach and perform semi-structured interviews with 26 participants belonging to 16 fact-checking organizations or fact-checking teams within publication houses. I began this study by

interviewing fact-checkers and editors—the stakeholder groups identified in previous works [49, 173, 362]. I discovered the existence of other stakeholder groups through these interviews and expanded the recruitment by reaching out to them using convenience [137] and snowball sampling [168]. The participants had diverse representations from 4 continents—North America, Europe, Asia, and Africa. I intentionally sampled participants widely across fact-checking teams, organizations, and countries to capture the practices and challenges emerging in this space. My work aims at uncovering the possible human and technological infrastructures supporting the fact-checking work in teams/organizations across the regions instead of capturing the variability of fact-checking process across regions.

My findings reveal the existence of six distinct stakeholder groups involved in the fact-checking process and the various roles performed by them. The identified stakeholder groups are: (1) *Editors* who are responsible for overseeing the fact-checking process, including planning what topics to target and ensuring the integrity of the fact-checks produced, (2) *External fact-checkers* who are responsible for monitoring the external world (social media platforms, presidential speeches, etc.), investigating dubious claims and writing fact-checks, (3) *In-house fact-checkers* who are responsible for fixing incorrect claims present in the news stories or articles produced internally in the media/news publication house, (4) *Investigators and researchers* who conduct in-depth investigation and data analysis of persistently circulating disinformation campaigns (e.g. investigating coordinated campaigns that used anti-Ruto hashtags[1] on Twitter to spread misinformation[2]), (5) *Social media managers* who distribute fact-checks across multiple social media platforms and strategize on ways to increase audience engagement with the fact-checks, and (6) *Advocators* who spearhead initiatives to improve policies around the availability of information and statistics in their countries to improve the quality of fact-checking. By studying the roles performed by the stakeholder groups (*human infrastructure*), I establish how fact-checking has evolved from a process to debunk individual pieces of misinformation (*short-term claims centric* fact-checking) to a multi-step long-term campaign involving research, policy, and advocacy work (*long-term advocacy centric* fact-checking). I find that stakeholder groups mediate their roles via different tools (*technological infrastructure*) ranging from third-party social media monitoring tools (e.g. BuzzSumo [80]), public databases, process management

[1]William Ruto is the current Deputy President of the Republic of Kenya. In May 2020, several anti-Ruto hashtags (e.g. #RutoMustGo, #RutoWantedToKillUhuru, etc.) began trending on Twitter in an attempt to discredit the Deputy President.

[2]https://investigate.africa/opt-report-post/

tools (e.g. Trello [392]), color coding schemes, to training and educational workshops. The interviews reveal that fact-checkers are skeptical of using fully automated AI-based tools. They desire algorithm explainability and the involvement of humans in the decision-making process as key values in the systems they would use. I also identify several technical, policy, and informational challenges. For example, there are limited tools to monitor and flag content on private messaging platforms and investigate false claims in videos and content in local regional languages. I also find that in some countries, information to investigate claims from government and civic bodies is either unavailable, difficult to obtain, or not updated periodically.

6.1.1 Research context: Human and Technological Infrastructures

The foundational work of studying infrastructure as a subject could be credited to Star and Ruhleder [367]. The authors consider infrastructure as "something that emerges for people in practice, connected to activities and structures" [367]. Rather than being a thing to use, Star and Ruhleder refer to infrastructure as a relational concept [367]. Scholars have since advocated for broadening the understanding of infrastructure by also including social practices, processes, and flow of information [340] and have called for investigating the complexities and particularities of infrastructures in practice [83, 226, 401]. In response to this call, several research studies in Computer Supported Cooperative Work (CSCW) and related fields, such as Human Computer Interaction (HCI) have examined the infrastructures—both human and technological—supporting the diverse socio-technical systems [131, 210, 252, 293] in various contexts such as, in health-care [141, 307, 375], e-governance [89], crisis situations [265], etc. For my study, I first focus on human infrastructure which Lee et al define as "organizations and actors that must be brought into alignment in order for work to be accomplished" [252]. Scholars have used this concept of human infrastructure to denote the human partnerships that are necessary for a successful socio-technical system [340]. Drawing on such scholarly work [131, 252, 340], I use the analytical lens of human infrastructure to "magnify the social" by rendering visibility to the stakeholder groups who collaborate to enable the fact-checking work. Highlighting the human infrastructure also allows us to focus on how collaboration and coordination are accomplished in socio-technical systems [252]. Sustaining online collaboration among various groups in an organization can be challenging [285]. Thus, within CSCW, a lot of attention has also been paid on examining collaborative [219, 244, 252] and coordination efforts [179, 263, 285, 412] in socio-technical systems, determining ways to foster

collaboration and coordination [155, 286, 339] and designing cooperative work tools [146, 185, 194, 227]. I complement these prior studies on by establishing fact-checking as a distributed problem that requires collaboration and coordination of the human infrastructure supporting the fact-checking process.

In a socio-technical system, human infrastructure does not exist in a vacuum [375]. It is intertwined with the technological infrastructure which is the software, hardware, and processes supporting the human actors in performing their roles [325, 375]. Scholars argue that technology and human actors are mutually constituting; one mediates the other [333, 375]. My work also borrows the concept of technological infrastructure to shed light on how the use of various tools facilitates the enactment of the stakeholder groups' roles in the fact-checking process.

6.2 Method

To better understand how fact-checking is practiced in real-world, I conducted semi-structured interviews with six stakeholder groups (N=26): (1) Editors (2) External fact-checkers (3) In-house fact-checkers (4) Investigators and researchers (5) Social media managers, and (6) Advocators. All interviews were conducted with the approval of the Institutional Review Board. I started by interviewing fact-checkers and editors employed in fact-checking organizations and publication houses. The qualitative analysis of the initial interviews (with P3, P4, and P5) as well as conversations with our contacts in the fact-checking organizations during the initial recruitment gave us a new perspective on fact-checking revealing the complex workflows that include several other stakeholder groups (apart from editors and fact-checkers) who work together to achieve the end goal of fact-checking. I then expanded my recruitment to interview people belonging to these other stakeholder groups.

6.2.1 Participant Sampling Technique

I adopted convenience [137] and snowball sampling [168] to recruit the subjects. First, I employed convenience sampling to identify fact-checkers via Twitter search. I sent out personal recruitment messages to those Twitter users whose Twitter bio revealed them to be fact-checkers and whose accounts allowed direct messaging. The second author had established collaboration with a few fact-checking organizations. I also reached out to individuals working in these organizations. Next, I used snowball

P#	Gender	Exp.(yrs)	P#	Gender	Exp.(yrs)	P#	Gender	Exp.(yrs)	P#	Gender	Exp.(yrs)
P1	Female	0.67	P8	Male	2	P15	Female	2	P22	Male	3
P2	Male	0.21	P9	Female	1	P16	Female	0.5	P23	Male	1
P3	Female	2.5	P10	Male	0.42	P17	Male	0.83	P24	Male	15
P4	Male	2.25	P11	Female	2	P18	Male	10	P25	Female	17
P5	Male	3	P12	Male	4	P19	Male	4	P26	Female	1
P6	Female	0.75	P13	Female	2.5	P20	Female	7			
P7	Female	0.92	P14	Male	0.5	P21	Male	0.75			

Table 6.1: Table showing list of participants with their gender and experience (in years) in their current role. Some participants have been associated with fact-checking work for a longer duration. I only report their experience in the current role in the organization.

sampling for recruitment. I requested the individuals who participated in the study to further connect us with other individuals belonging to different stakeholder groups in their or other organizations. I interviewed a total of 26 participants from 16 fact-checking teams/organizations including a Pulitzer Prize-winning editor and journalist. Tables 6.1 and 6.2 list the participants' demographics, experience in their current role, stakeholder groups that I studied, participating organizations, and the continents I covered by interviewing participants based in those continents. Note that the names of the roles of various stakeholder groups were not always self-reported (with the exception of fact-checkers, news editors, and copy editors), but were rather found during qualitative analysis. Therefore, these roles might not match with participants' designation in the fact-checking team or organization. For example, the professional designations of two participants performing advocacy work in their respective fact-checking organizations were *Head of Public Policy & Institutional Development* and *Partnerships manager*. Several of the participants' roles were fluid and overlapping, i.e. in some fact-checking teams/organizations, at times a single person enacted several roles. For example, participant P12 performs both editorial and advocacy work. In some cases, a participant provided us with insights about more than one role since they either led or managed the entire fact-checking team or worked closely with the other stakeholder groups and thus, were aware of the various roles involved in the fact-checking pipeline. For example, I interviewed an advocator working at Meedan, an organization that provides institutional and programmatic support to partner organizations doing fact-checking work. Their job at Meedan allows them to work closely with people performing various roles at the fact-checking organizations. Thus, the participant was able to describe the responsibilities and tasks conducted by various stakeholder groups such as fact-checkers, advocators and investigators.

Stakeholder group	Organization	Continent
1) News desk editors and copy editors 2) External fact-checkers 3) Social media managers 4) Investigators and researchers 5) Advocators 6) In house fact-checkers	Pesacheck [311], Meedan [271], First Check [250], Full Fact [139], The African Network of Centers for Investigative Reporting's Investigative Lab [33], AFP [43], Africa Check [91], The New Republic [326], The Quint [322], Al Jazeera [214], The Washington Post [381], DPA [13], Maldita [15], India Today [14], Der Spiegel [12], Fine Tip Research & Editing, Freelance	Africa, Asia, Europe, North America

Table 6.2: Table showing the stakeholder groups identified in the study, the participating organizations, and the continents I covered through the interviews. In the organization column, freelance refers to no association with a particular fact-checking organization/team. I aggregated the roles of stakeholders and their association with fact-checking organization/team to ensure anonymity as in some cases knowledge of network affiliation and role could potentially reveal the identities of a few participants. Note that the participants that I interviewed sometimes provided insights about more than one role.

6.2.2 Interview Protocol and Data Analysis

All interviews were conducted between November 2020 to September 2021. I designed a generic semi-structured interview script for the study that contained a set of broad questions about participants and their organization's role. Based on participants' responses to these questions, I inquired them about the specific details of their roles. Thus, recruiting and interviewing different stakeholder groups did not require us to file any changes with my university's Institutional Review Board. I first asked participants to describe their role within their organization and the function of the organization itself. I encouraged participants to share their screens and describe various aspects of their work using real-world examples. I also asked the participants to demonstrate the tools they use wherever applicable. I probed them about the role of technology in their day-to-day work and how the affordances provided by online platforms facilitate or impede their work. To get insights about how fact-checking is practiced in the participant's team/organization, I asked them to describe all the steps involved in the fact-checking pipeline. I inquired about the other stakeholder groups working in their team/organization who also contribute towards the fact-checking process and how these various groups collaborate with each other. I also discussed various challenges participants face in their job. The interviews lasted between 60 to 125 minutes and averaged over 90 minutes. All interviews were recorded via Zoom or Google Meet. The first author transcribed all the video and audio recordings. Then, two authors independently went through the transcripts and observation notes taken

during the interviews and thematically analyzed them using a mixture of deductive and inductive coding schemes [74]. First, both authors conducted a deductive scan where the transcripts were coded for categories: fact-checking process, stakeholder groups, participant's role, decision making, use of tools, collaboration within stakeholder groups, and challenges faced by the participant in performing their role. Then within each deductive code, inductive coding was conducted. Both authors read the transcripts multiple times to determine the codes. Then, the two authors compared and contrasted their codes with each other to refine the codes and resolve the inconsistencies. After several rounds of discussions, both authors converged on a final set of themes. I present these themes with respect to the research questions in the following sections.

6.3 Types of Fact-checking: Short-term Claims and Long-term Advocacy

This study aims to identify the infrastructures—both human and technological—supporting the fact-checking work. I identify and present the human infrastructure by elucidating the role of six stakeholder groups that need to come in alignment to accomplish fact-checking. The six stakeholder groups are editors (news desk and copy editors), external fact-checkers, in-house fact-checkers, social media managers, investigators and researchers, and advocators. I show how these stakeholder groups' roles are supported by technological infrastructure. Through the study of the fact-checking infrastructures, I establish that fact-checking exists as both *short-term claims centric* and *long-term advocacy centric* fact-checking. In this section, I first provide an overview of the two types of fact-checking before deep diving into the infrastructures supporting them in Section 6.4 and Section 6.5. Figure 6.1 provides an overview of the fact-checking ecosystem including the stakeholder groups, their roles, and various tools that support them in performing their roles.

6.3.1 Short-term Claims Centric Fact-checking

Short-term claims centric fact-checking aims at informing the public by debunking misleading claims circulating on online platforms. It begins with fact-checkers continuously monitoring the online spaces for potentially misleading content. They identify the exact claim(s) that they want to fact-check and make a pitch to the editorial team

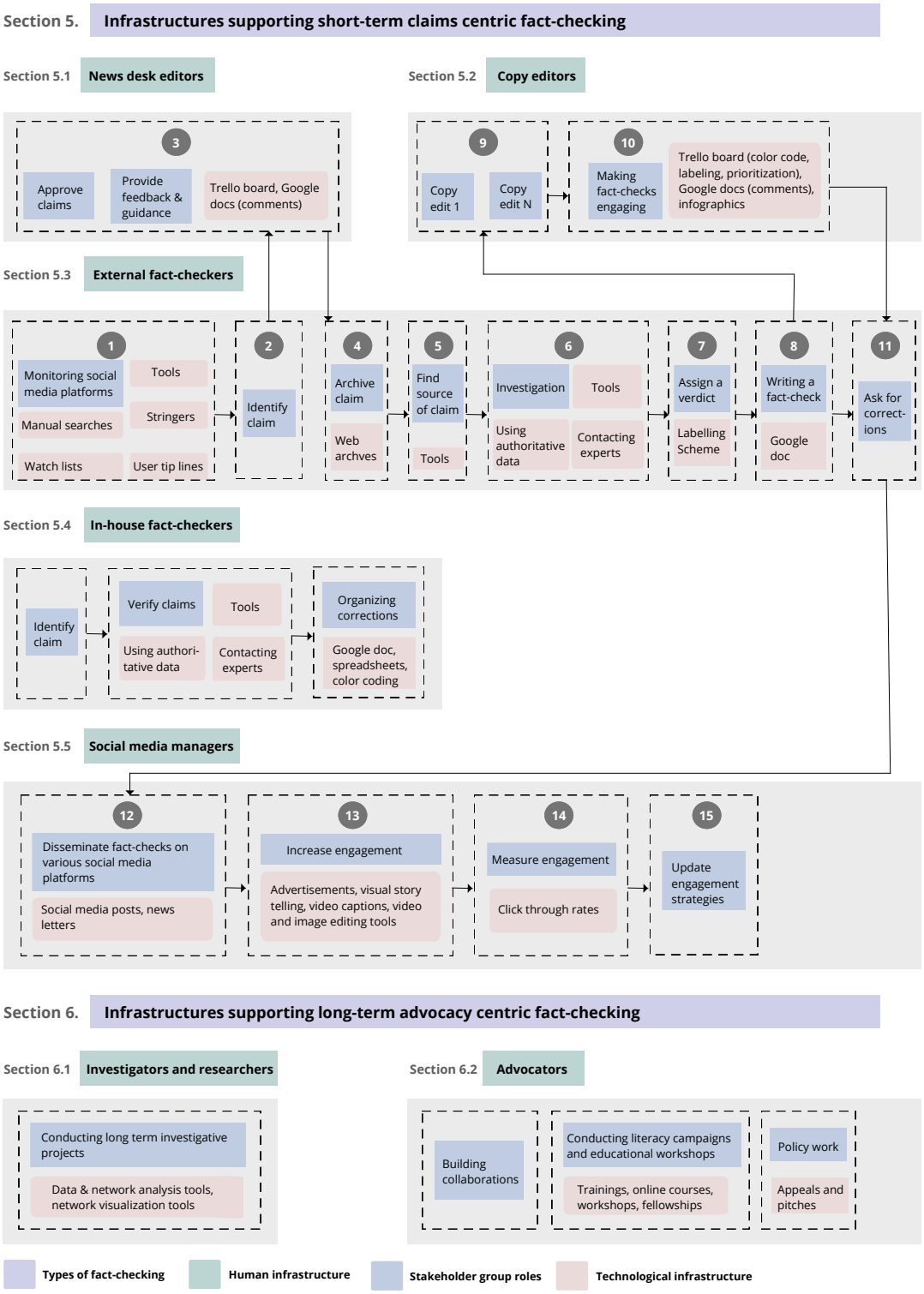


Figure 6.1: Figure presenting the ecosystem of fact-checking, the whole or part of which could exist in a fact-checking organization or a news publication house. ▢ indicates the two types of fact-checking (*short-term claims centric* and *long-term advocacy centric* fact-checking) introduced in the study, ▢ presents the stakeholder groups involved in the fact-checking process (human infrastructure), ▢ shows work done by the stakeholder groups as part of their role, and ▢ specifies the tools stakeholders use to mediate their roles (technological infrastructure). The numbers indicate the sequence in which various roles are performed.

about how they plan to debunk that claim (Section 6.4.3). After gaining the editor's approval (Section 6.4.1), they archive the content, find the source of the claim, and investigate the claim by using online tools, consulting experts and employing authoritative publicly available evidence (e,g, statistics on unemployment, census data, etc.). Based on their investigation, fact-checkers assign a label indicating the veracity of the claim and then write a report declaring all the sources they gathered (Section 6.4.3). The written report goes through a rigorous copy editing pipeline to ensure the integrity of the fact-check (Section 6.4.2). If the claim is false, fact-checking organizations reach out to the person/organization who made the claim for correction(s) (Section 6.4.3). Finally, the social media engagement team publishes the fact-check story/document on the social media pages of the organization and adopts several strategies to increase the audience's engagement with the published fact-check (Section 6.4.5). My work also examines the role of in-house fact-checkers doing *short-term claims centric* fact-checking in news and media publication houses. In-house fact-checkers assist reporters/journalists in verifying and validating their articles by ensuring that the facts and quotes present in the article are correct and backed by authoritative sources (Section 6.4.4).

6.3.2 Long-term Advocacy Centric Fact-checking

> "Fact checking is more than publishing fact checks. In order to change the information ecosystem.., we need to spot patterns and try and do something about those patterns.. We try to influence policymakers, information producers, and media to raise their standards and improve the quality of information and public debate. [Our role is] more akin to a campaigning organization. " - P9

I find that the work of the fact-checking organizations is more than a one-off engagement with misleading claims and fact-checks (*short-term claims centric* fact-checking). Most organizations also perform *long-term advocacy centric* fact-checking. They run several investigative projects to study the misinformation ecosystem in their country (Section 6.5.1). They are also actively involved in advocacy and policy work where they try to influence civic bodies and policymakers to improve the quality of data, organize workshops to train organizations and journalists to do fact-checking, and work towards forming coalitions among various fact-checking organizations and internet companies (Section 6.5.2). To demystify the fact-checking process, in the following sections, I present in detail the roles performed by the stakeholder groups involved in both types of fact-checking along with the collaborations occurring in the

process. For each stakeholder group, I also discuss how technological infrastructure supports the enactment of their roles.

6.4 Infrastructures Supporting Short-term Claims Centric Fact-checking

Short-term claims centric fact-checking is supported by five stakeholder groups—news desk editors, copy editors, external and in-house fact-checkers, and social media managers. I present the roles played, activities performed, decisions made, and tools used by them.

6.4.1 News Desk Editors—Approving Claims and Guiding Fact-checkers

News desk editors are one of the most critical stakeholder groups supporting the *short-term claims centric* fact-checking. They decide what their team/organization is going to fact-check. They approve or reject the claims pitched by the fact-checkers and guide them in their work.

6.4.1.1 Approving claims to fact-check

News desk editors are looking for newsworthy claims that impact a lot of people. The first criteria for approving a claim is the popularity and reach of the person making the claim since it increases the chance of the claim spreading far and wide ("*it should be newsworthy, said by an important person,... if a popular public figure makes a claim, a lot of people are likely to be exposed to that claim, given the bully pulpit that public figures have"—P18*). The second criterion includes a number of people likely to be misled based on the context surrounding the claim. For example, a claim touching upon a communal angle is likely to impact a lot of people in a country that has "*many illiterate people..who process information based on their communal experience* (P12)". Third, news editors approve content that is gaining a lot of traction on social media platforms by accumulating engagement in the form of likes, comments, and shares. Content that has received less attention from the public is not fact-checked from the fear of amplifying the false information by inadvertently bringing attention to the false claims (*"Are people believing this or are they taking it just as a joke? Does it just have one share,*

meaning if I fact checks it, it will just amplify the fake news and not really give the correct information"—P11). Fourth, opinion pieces and claims that cannot be verified using sources are not fact-checked ("*we check things that have facts, that can be verified using records and information. We don't want a situation where we see this person says this and this person says this, that will just be hearsay"—P3*). Fifth, fact-checkers and news desk editors consider several stages of harm that false information is likely to cause—physical, mental, social, or emotional and prioritize the claims that are likely going to cause maximum damage to the public by affecting their health and well-being.

> "We have stages of harm that you have to check. Is it causing physical harm? Is it causing mental harm? Is it causing someone to lose social standing? How much effect will it have to the person.. if I don't fact-check the story? .. We do the ones that actually have greater harm, we give them priority.. We have to get it out before most people see it to reduce the harm that it's causing." - P3

6.4.1.2 Guiding fact-checkers

The job of news editors do not end after approving a claim. They also guide and help fact-checkers in "*gathering sources and evidence (P11)*" for verifying stories, "*understanding concepts, finding best available data [for research],..and connecting with the experts given [editors'] long experience in the media ecosystem (P12)*".

6.4.2 Copy Editors—Ensuring Quality of the Fact-checks

Copy editors do quality control of the fact-check story/document through multiple iterative copy editing cycles. A fact-check story or document is a report written by fact-checkers containing the claim investigated, sources used for investigation, and verdict indicating the veracity of the claim. This stakeholder group acts like the *first readers* who determine whether fact-checkers have accurately interpreted the claim in the fact-check, used multiple primary sources as evidence for the investigation, provided working links to the sources used for investigation, and presented the evidence in such a way that it leads to a logical and correct conclusion. They also work towards making the fact-checks engaging by checking the phrasing and grammatical errors. I discuss the tasks performed by copy editors in detail below.

6.4.2.1 Performing copy editing cycles

Through the interviews I realized that a fact-check goes through two to three copy-editing stages. Each stage is supervised by a different copy editor to ensure higher quality. The written fact-check is provided to the copy editor in a shared document (e.g. Google doc) where they leave comments to provide feedback. At the first copy edit stage, copy editors check the central premise of the fact-check.

> "When I get a fact-check, I read through it three times to understand it before I make any change on it, before I ask any questions. I look for the claim and the debunk. Does it really hold? If there are questions about the debunk then we send it back to the news desk.. At that stage, the fact checker will pick it up and go back and try to sort out any queries that have been raised." - P1

The second copy-edit stage focuses on refining the language and flow of the fact-check. The final copy-edit stage focuses on making the fact-check more engaging and interesting to read. I discuss this aspect briefly.

6.4.2.2 Making fact-checks more engaging

Copy editors try to keep the fact-check short, clear, crisp, and interesting. They ensure that the fact-checks are written in a language that is understood by laymen. P1 spoke most candidly about the engagement aspect of the editorial process. They revealed that they often collaborate with social media managers to get feedback on the engagement aspect of the fact-checks. For example, ensuring that the country relevant to the fact-check is present "*in the title or in the blurb (P1)*" so that people could quickly determine if it's of interest to them, or making certain that the verdict on the claim is placed high up in the fact-check to get more attention from the public. Copy editors along with social media managers also suggest the addition of info-graphics (engaging visuals, imagery, tables, charts, etc.) to fact-checks in order to attract the eyes of the readers. The info-graphics are added to quickly communicate "*complex information in a visual manner* (P2)". They are usually added in long-form fact-checks—the ones that debunk multiple claims and delve into the subject of the claim in greater depth.

Copy editors also ensure that the fact-check contains terms that people are most likely to search online. P1, P9, P12, and P18 talked about the claim-review schema used by fact-checking organizations to allow internet companies like Google to index their fact-checks [153]. These participants believe that using popular search terms in fact-check helps increase its visibility since search engines then rank it higher in the search

results. I briefly touch upon this collaboration between fact-checking organizations and internet companies later in Section 6.5.2.

6.4.2.3 Use of collaborative systems, labeling, and color coding schemes

In most of the organizations that I interviewed, both news desk and copy editors use Google Docs to comment and provide feedback on the fact-checks. In addition, fact-checking organizations like Pesacheck, Africa Check, etc. also use an open-source project management and collaboration tool called Trello [392]. The tool provides a dashboard with a series of columns that contain cards. The columns are named so that they denote the current stage of the fact-check. For example, P2 showed the Trello board of their organization which had several stages such as *complete fact checks, copy edit 1, copy edit 2, copy edit 3, Q/A & final review, final review done, published live, and amplification done*. Trello cards are the basic functioning unit in the tool. The cards hold various information about the fact-checks including due dates, conversations, attachments, etc. The tool allows editors to add colors and labels to the cards. Copy editors use labels and color coding schemes for different purposes. For example, P2 uses labels for prioritizing the fact checks—("*Earlier this year there was a lot of interest in COVID, so I would label fact-check as COVID because we were prioritizing those*). P10 uses colors to indicate tasks allocated to people ("*It's easier for me to see what people are working on.. If its research, it will be purple, if its documentation, it will be green.*)."

6.4.3 External Fact-checkers—Monitoring, Investigating and Publishing Fact-checks

External fact-checkers are the most evident stakeholder group supporting the fact-checking process. Their role is to continuously monitor the external world for potentially false claims, investigate them for veracity, assign a verdict, and publish fact-checks. I discuss these roles below. I also specify the technological infrastructures supporting the roles along with the description of the roles.

6.4.3.1 Monitoring online spaces

Monitoring content is one of the most tedious steps in the fact-checking workflow. Fact-checkers monitor content reactively (in response to user tips), preemptively (before events like elections, presidential speeches, etc.), and in real-time (tracking current affairs via trends and listening to conversations in real-time). They monitor the content

via several tools and artifacts. First, they rely on user reports and tiplines which are useful ways to access misleading content circulating in private groups and WhatsApp that are otherwise hard to access.

> "We have our WhatsApp tipline and emails.. Public who wants to verify a particular piece of information send it on WhatsApp to us.. We verify those queries at our end and then send them the replies with the fact-check story if possible. If we don't have fact check story, then we send people whatever information we have on the query. " - P14

Second, fact-checkers create watch lists and track social media accounts, groups, pages, and websites of repeat offenders–those who posted misinformative content multiple times in the past (*"We have a database where we've tracked all accounts spreading misinformation.. We go back and check these accounts, see what they've posted on their website and personal account."—P3*).

Third, fact-checkers rely on manual searches. They follow current events via news or Twitter trends to get updated about topics that people are talking about and track them on all online platforms. Creation of relevant search queries to track these topics is mostly a tedious "*hit-n-trial* (P7)" method. Searching for a query can give millions of results on search platforms. Therefore, fact-checkers rely on search query syntax to reduce the number of search results.

> "I do not want the news content. I want user generated content. So one of the simplest tools is to write minus news (search_query -news) so that it cancels out the major news content." - P7

> "I use keywords like "intitle" [on Google search]. [intitle:coronajihad site:facebook.com].. is showing me every search term, every post on Facebook, which has the title "coronajihad"." - P8

Fourth, fact-checkers use several tools to track content on the internet. For example, they use CrowdTangle [107] to track Facebook's public pages and groups. Organizations that have partnered with Facebook have access to a Facebook proprietary tool colloquially known as the "Facebook Queue". The tool aggregates potentially misleading content that is accumulating engagement on the platform. In some of the organizations, fact-checkers also rely on several third-party tools (e.g. Social searcher [352], Influencer [207], BuzzSumo [80] etc.) to search and filter content on platforms since they provide them varied search filter options that the original interface of the

social media platform lacks. For example, Facebook search allows one to filter content by year and not by months and dates. Lastly, fact-checkers also rely on a network of stringers—"reporters who work for a publication or news agency on a part-time basis" [414]—who inform them about the misinformation "*circulating in their region..[and] language* (P11)."

6.4.3.2 Making a decision to fact-check and extracting claim(s)

Once fact-checkers have identified potentially false content on the internet, they identify claim(s) in the content to verify. The fact-checkers can decide to do a short-form or a long-form fact-check depending on the number of claims they identify in the content. For their partnership with Facebook, the fact-checkers only do short-form fact-checks where they identify one claim from the body of the content. In long-form fact-checks, fact-checkers will debunk multiple claims present in the content.

> "The reason we do [short-forms] is to make sure we are clear, we're not confusing the audience. This is largely inline with the Facebook partnership where we mainly focus on one claim." - P2

Next, fact-checkers prepare a pitch to convince news desk editors why the content needs to be fact-checked along with a plan of how they would accumulate proofs to debunk the claim. Once the claim is approved, they archive the content using online websites like archiveis [52], wayback machine [51], etc. Many times people and organizations delete the false claims made by them once they are fact-checked. Thus, archiving becomes essential to prove that the content existed.

> "Once you identify the claim you archive it because what purveyors of misinformation do mostly, they delete it. So once they delete it you're unable to read that particular claim." - P11

After archiving, fact-checkers find the original source of the claim–"*who shared [the claim], context with which it (the claim) was originally shared* (P14)", without which the fact-checker would not have a complete picture of the context in which the content was originally shared.

6.4.3.3 Researching

Once a claim is identified, fact-checkers collect multiple primary sources to prove or disprove the claim. They use three ways of gathering sources. First, they use quotes

from experts such as doctors, physicians, meteorologists, and academics. Second, they use public authoritative sources like government databases (e.g. Kenya National Bureau of Statistics), mainstream news sources (e.g. CNN), and peer-reviewed scientific research papers and journals. Third, they rely on several tools. For example, fact-checkers use image and video verification tools such as InVid [209] followed by a reverse image search on search engines to collect metadata and a digital trail of the image/video in order to determine its authenticity. Fact-checkers from First Check, Pesacheck, and Africa Check also reported that they depend on third-party tools such as whois [418], Spoonbill [366], edit history functionality in Facebook, etc. to determine the veracity of a website or post. Through my interviews, I realized the important role played by comments in fact-checking. Comments contain useful clues that help in investigating the claims.

> "What we do first is read comments before checking. We found in comments that one woman said this is not the entire video, here is the link to entire video. So that's what led us to entire video. [In comments] we see some clues, how to look for what really happened. " - P16

6.4.3.4 Assigning veracity label and publishing fact-check

After the investigation, fact-checkers assign a label to the claim that reflects its veracity. Through the interviews, I realized fact-checking organizations use a range of labeling conventions, for example, 5 point scale ranging from completely false to true, a four-point Pinocchio scale [314], etc. There is no commonly accepted standard for labeling misinformation. By publishing these details, the fact-checker takes the reader through the entire investigation "*so that the readers can replicate [the process] themselves* (P18)".

6.4.4 In-house Fact-checkers—Gathering Sources and Verifying Claims

In-house fact-checkers are employed by publication houses to fact-check the stories produced by the journalists before they are published and disseminated. They receive the script or the news story from the journalist along with all source material that they used while researching and writing the piece. The in-house fact-checker then verifies every claim present in the story and delivers a modified story along with a list of proposed changes. Unlike the job of external fact-checkers, the job of this stakeholder group is to fix or remove incorrect claims without publicizing or calling attention to the

inaccuracies [172]. The need for in-house fact-checkers arises in a publication house not only to ensure that the stories published are reporting accurate facts to readers but also to *"protect the publishing house from any liability or future lawsuit (P13)"*. I discuss the roles of this stakeholder group below.

6.4.4.1 Identifying and verifying claims

In-house fact-checkers verify every line present in the story. Each line usually has more than one fact to be checked, from how a proper noun is spelled, and grammatical idiosyncrasies to every phrase making a claim. Journalist's opinions and arguments, and quotes from a reputable expert are the only phrases in the content that are not verified. However, for opinions, the in-house fact-checkers check the context surrounding the argument to ensure it is "*right and mainstream (P13)*".

All the claims that are vague and without proper sources backing them are modified or removed from the text.

> "Vague and broad facts which I.. try to get people to remove.. I recently had a whole debate with someone about a line that said America is more divided today than ever. I was like, how do you check that line.. what are the sources for that kind of statement? [It] is just so broad." - P15

In-house fact-checkers employ a myriad of techniques— top-down approach, prioritization, and batch processing—to identify and verify claims. The majority of the in-house fact-checkers I interviewed first perform a top-down linear scan of the document to get the most central ideas "*which if were incorrect, the entire piece (article) would be called into question*" (P13). Second, they prioritize the claims for verification. Different fact-checkers prioritize claims differently. For example, P13, P15, and P25 work with organizations that have strict timelines for publications. Thus, they first prioritize the claims that will take the maximum time to investigate. On the other hand, P20 does not have strict deadlines and prioritizes claims that they are certain are false.

> "[I prioritize] a claim that I know is going to take me a while to nail down. So for example, if there's a claim ... about someone committing a crime and I need to file a FOIA request[3] or go through public records or order a document from sort of government agency, I want to do that as early as possible so that I get myself as much time. You know, basically, any claim that relies on other actors in order to meet to verify." - P13

[3]https://www.foia.gov/how-to.html

> "When I prioritize I start with the things that I know are wrong, and then I look at the things I think are right. And then I check all dates and all names. " - P20

P15 revealed that they *"process the claims in batches"*. Everyday they work on a batch of claims and send inquiries and suggestions regarding those claims to the journalist. This technique gives journalists ample time to respond and does not inundate them with several queries towards the very end of the schedule.

6.4.4.2 Gathering sources for verification

In-house fact-checkers use "*primary reputable sources (P15*)" to verify claims. The journalist is expected to give the primary sources that they used while doing research and writing the article. However, if the sources are missing, in-house fact-checkers look for primary documentation (like death certificate, house deed, etc.), mainstream news sources (like CNN, New York Times, etc.), academic peer-reviewed journals, and relevant experts to verify the claims. In case the source used by the journalist is not reliable, the usual journalistic practice is to gather a total of three to four sources to back up the claim. In-house fact-checkers heavily rely on Google search to hunt for sources. Some of the in-house fact-checkers also use Nexis search [17]—a paid service—that gives news articles, blogs, and legal documents as search results. They find the latter to be more effective for searching news and documents.

6.4.4.3 Organizing corrections

After identifying and verifying claims, in-house fact-checkers organize and communicate the list of questions and suggested changes to the journalist. According to in-house fact-checkers, there is no standard convention for organizing corrections. However, I found that all fact-checkers use color coding schemes for the organization but in different ways. For example, P13 highlights verifiable and unverifiable claims in the Google doc with different colors and leaves comments containing information about the sources. P15 copies each claim from the doc into a separate row in a spreadsheet and then uses color coding to indicate the state of the claim (pending, verified, incorrect, etc).

6.4.5 Social Media Managers—Disseminating Fact-checks, Increasing Engagement

Social media managers are responsible for leading all social media initiatives including publishing and disseminating fact-checks on their organization's social media handles. They determine how to make fact-checks more appealing so that they attract people's eyes. They also measure the engagement received by the posted fact-check(s) and based on the feedback continuously update their amplification strategies to attract more audience to engage with the fact-checks.

I discuss these tasks below.

6.4.5.1 Disseminating fact-checks

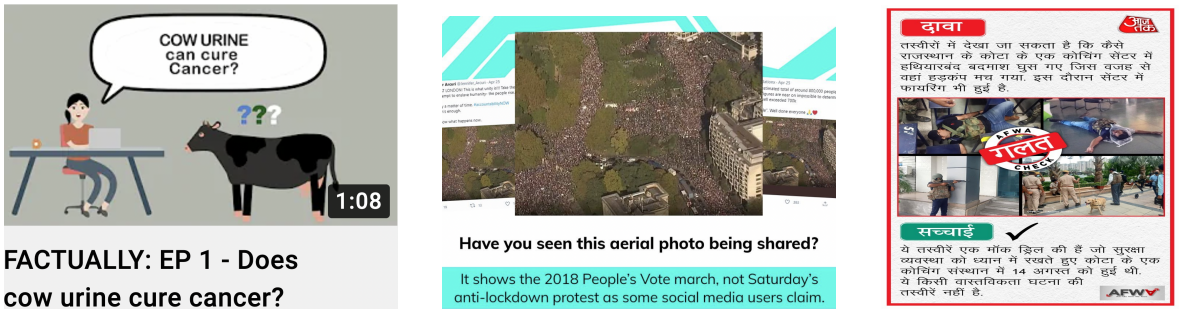
The social media manager's primary responsibility is to post fact-checks and educational tip sheets produced by fact-checking organizations on all major social media platforms. Few organizations also have a "*WhatsApp number where they broadcast weekly newsletters containing the top fact-checks of the week (P12)*". Social media managers want to make their fact-checks accessible to people with disabilities. Since Twitter did not have a captioning feature with its audio tweets, P17 usually posts fact-checking videos with captions. They use tools like Kapwing [26] to add captions to the videos.

> "We wanted to see how can we reach, for example, blind and deaf audiences. The audio tweets did not have captioning, they did not have accessibility features for disabled audiences. And so this is why we decided to use the video instead because videos allow us to caption and so people who can't hear the video can still read the information. " - P17

6.4.5.2 Adopting strategies to increase engagement

Social media managers adopt several innovative ways to increase their content's reach inorganically (via ads) and organically (via visual storytelling). I discuss some of the strategies.

- *Running advertisements:* A few fact-checking organizations that have partnered with Facebook receive free advertisement credits to promote their fact-checks on the platform. P17 delved deep into the ad usage process. They advertise fact-checks that are not bounded by time since they could be promoted via ads for an extended period. To select the audience for ad targeting, they look at three attributes, namely, the audience's education, relevance to the fact-check, and



(a) (b) (c)

Figure 6.2: (a) A short YouTube video explaining a fact-check using comic like visuals (b) An Instagram post containing a fact-check (c) A "postcard" containing fact-check in Hindi language to be shared on mediums like WhatsApp. The single image contains the false-claim and the debunk.

interests. They target audiences who have "*graduated from a university*" (education), live in "*countries that are most relevant to the fact-check*" being promoted (relevance), and are "*interested in news, advocacy and community issues*" (interests).

- *Visual storytelling:* Social media managers find visual storytelling to be a very effective way of getting people to engage with fact-checking content. They've found that many people prefer "*watching their content over reading it (P18)*". Therefore, they convert fact-checks into a visual narrative (images or videos) before posting them on social media platforms. To create the visual content, these stakeholder groups rely on video and image editing tools. For example, social media managers in a few of the organizations that I interviewed use multimedia editing tools such as Adobe Illustrator [25], Photoshop [28], etc.
  Social media managers also create comic strips where a cartoon walks readers through the fact-checks ("*I myself had started this small comic strip thing to get more engagement.. it became quite popular and people were liking it, and sharing it*"—*P7*). To engage with the local non-English speaking audience, the comic strips, images and videos are also converted to regional languages.

> "We translate [comics] into Swahili because there's been this type of content, .. being done in like mainstream languages like English, French, Portuguese, But some of the more widely spoken local languages aren't really a priority. That was a gap that we identified. " - P5

Figure 6.2 contains three examples of fact-checks leveraging visual storytelling techniques.

6.4.5.3 Measuring engagement and updating strategies

Measuring social media engagement is important to determine what content is gaining more traction and which engagement strategy is working. P17 informed us how they track click-through rates to determine how many people are engaging with the content by clicking on the links posted by them.

> "[Bitly] can tell how many people clicked on a particular link, and we can track that.. [It allowed us to] analyze how people engage with these links, how many clicks come from the newsletter, as opposed to Facebook or Twitter, or Instagram? And then that allows us to know whether we need to change how we present the information. " - P17

The engagement statistics help social media teams to update their engagement strategies. For example, P17 described how they changed the number of stories published in their weekly WhatsApp newsletter from five to three after realizing that the public only clicked on the top URLs.

> "Earlier editions [of WhatsApp newsletter] had up to 5 stories per edition. And we were seeing that people were only clicking the top 2 or 3 links, and were ignoring the rest. So a decision was made to reduce the number of stories that we feature from 5 to 3 to make it shorter." - P17

6.5 Infrastructures Supporting Long-term Advocacy Centric Fact-checking

Long-term advocacy centric fact-checking aims at improving the information landscape by conducting research about various aspects of online disinformation, influencing policies surrounding the availability and quality of data and statistics, conducting educational training for aspiring fact-checkers, organizing literacy campaigns for the general public and forming coalitions with various fact-checking organizations and internet companies. This type of fact-checking is supported by two stakeholder groups—*investigators and researchers* who conduct long-term investigative projects that include data and network analysis, and *advocators* who are involved in policy and advocacy work. In this section, I elaborate on the role of these stakeholder groups.

6.5.1 Investigators and Researchers—Conducting In-depth Research and Investigation

Few fact-checking organizations (e.g. Full Fact, Code for Africa, etc.) have a separate team of investigators and researchers. Unlike fact-checkers who engage with individual pieces of misinformation, this stakeholder group conducts an in-depth investigation of persistently circulating misinformation and disinformation campaigns via data and network analysis.

> "Part of my work at the moment is creating or developing a framework for misinformation crises. So, as opposed to individual pieces of misinformation.. looking at when responses need to go over and above the day-to-day..everybody who works in the kind of anti-misinformation space gets together and they normally introduce new responses and new policies to manage that." - P9

> "If.. we are seeing a certain type of misinformation occurring.. every single day, debunking each individual one will not help. So what they [fact-checking team] do is now they refer that case to us. And then..[we do an] in-depth investigation to try and see where is this narrative originating from.. The data analytics team is the one that does the sifting through the data sets that we obtained from social media. The forensic team does profiling of key accounts..that we identified." - P10

6.5.1.1 Conducting long-term investigative projects

Investigators and researchers undertake several investigative projects such as verifying the backfire effect[4] of fact-checking (when a claim aligns with a person's beliefs, proving that it is wrong will make them believe it more strongly), studying long term effects of conspiracy theories, examining public engagement with political news, determining how to communicate fact checks effectively, etc.

> "You've probably heard about the backfire effect which is a kind of mythical idea that fact-checking.. does more damage than good. I think the original research was repeated and the same effect wasn't found.. We also did a project recently where we looked at.. how conspiracy theories affect people's beliefs in the long term.. who believes and shares misinformation, how to communicate your fact checks [while] presenting them to the audience." - P9

[4]https://fullfact.org/blog/2019/mar/does-backfire-effect-exist/

> "Projects include doing investigations into Russian disinformation or Russian influence in African countries.. investigating Chinese influence operations into .. African countries,.. human trafficking in four South African countries." - P10

> "There's been a lot of research that was done around.. how to design a fact check to make it more engaging, how do you phrase a headline,.. how much of this information can you put into like a video. What is the ideal video length? How to get more people to interact with them and then how to try to make them much more long-lasting and persistent in people's memory." - P5

These stakeholder groups analyze misinformative content that is going viral on social media platforms and then trace it back to the social media accounts that started sharing it. They also conduct qualitative research by conducting surveys.

> " We have like a survey and we've been trying to find what kind of narratives against immigrants are more popular here in Spain. We're trying to prepare another big survey about how fact-checking works in Spain and what kind of debunk works better for us. " - P21

The results of the investigative projects are released as dossiers (e.g. [16], [18]). The dossier provides a background of the issue investigated, data collection and analysis method(s), results of investigation, and conclusion. For example, in [18], investigators study the misinformation influence operations that occurred in Uganda before the January 2021 elections. The dossier first briefly describes the political landscape in Uganda and provides examples of misinformative tweets that acted as the starting point of investigation. These tweets leveraged trending hashtags such as #StopHooliganism, #UgandaIsBleeding [5], etc. to spread false narratives against the opposition party using past events from other countries. The dossier then deep dives into the methods used to collect and analyze relevant tweets and finally attributes the influence operation to the supporters of National Resistance Movement[6].

Investigators and researchers use tools like Python libraries for data analysis and Gephi [22] for network visualization in addition to the tools used by fact-checkers.

> "We were actually using an open source tool called tweepy to collect tweets from Twitter.. The network analysis is [done] using a tool called gephi [which] will

[5]In November 2020, Robert Kyagulanyi, a presidential candidate in Uganda was arrested on two separate occasions. The protests that broke out after his arrests were documented in Twitter posts containing hashtags such as #StopHooliganism, #UgandaIsBleeding, etc.

[6]https://en.wikipedia.org/wiki/National_Resistance_Movement

be able to show relationships between tweets using the retweet function like which are the accounts that have been highly retweeted, which are the influential accounts within the network. " - P10

6.5.2 Advocators—Influencing Policy, Building Coalitions, Conducting Educational Workshops and Literacy Campaigns

Several of the study participants revealed that fact-checking doesn't stop with the generation and distribution of fact-checks and is much more like a sustained campaign. It also includes influencing policymakers and information producers to improve the quality of information (and in turn the quality of fact-checking), building coalitions with other fact-checking organizations, social media companies, and journalistic organizations as well as providing fact-checking training to newsrooms and organizations.

Such initiatives are led by advocators. This stakeholder group identifies and realizes several ways to improve the *short-term claims centric* fact-checking process. They steward multiple outreach programs, policy initiatives, and advocacy projects locally and globally. All the initiatives started by these stakeholder groups could be considered actions that are performed via technological and informational infrastructures such as workshops, appeals, and training programs. I present the tasks performed by this stakeholder group below.

6.5.2.1 Creating new generation of fact-checkers and fact-checking organizations

The advocators conduct training, workshops and fellowship programs for people and organizations all over the world, teaching them nitty-gritty details of fact-checking along with how to setup and operate a fact-checking organization of their own.

"Ethiopia.. [is] a country where press freedom is very limited, online false information often leads to offline.. And most of the media there are state-controlled. So we were conducting training on how they can set up fact-checking desks and try to be independent." - P2

"In Germany, we only got two IFCN signatories. And that's not enough. So we try to convince traditional media outlets in the regions..to start with fact-checking.. We are training them. [We have created] a community of fact-checkers..and more than 600 journalists and we are doing training, encouraging them to start with fact checks." - P19

Advocators conduct training through webinars or online platforms. Some organizations have set up their online learning platforms where they provide video tutorials to fact-checkers (e.g. [125]), while others have partnered with academic universities to conduct educational training.

> "The advocacy campaigns.. on fact-checking, we conduct them virtually. We have partnerships with Kenyan institutions of higher learning, such as Daystar and Aga Khan university, which allows us to conduct webinars on fact-checking. Those sessions are attended by media professors and their students. We mainly use Jitsi, Slack, and Google Meet to do the training. " - P2

> "We are proud of the digital e-learning platform we have built at DPA. we have a lot of videos here, where we explain how to work with the Internet Archive. Those are webinars [where we] train maximum of 15 people, zoom webinars mean it's live training. " - P19

6.5.2.2 Pitching the importance of evidence-based decision-making to policymakers

The advocators are also actively trying to reach their country's policymakers and civil society organizations, informing them about their work and bidding the importance of facts and evidence-based decision-making. For example, P12 attended parliamentary researchers' conference in Kenya where they pitched the importance of facts in informing the country's policies and intervention programs. Similarly, P9's organization tried to get parliamentary support to attribute electoral imprints to election campaigners during referendum and elections in the United Kingdom.

> "Civil society organizations generate data that then is used by government to put in place policy intervention, say poverty eradication.. interventions in healthcare. So we just want them to understand that as you're doing this, you also have to check your facts, don't just rely on a news story or rely on a document or rely on a statement by a public politician to define your problem.. And so I.. tell them.. their job is to tell the leaders what the data says. " - P12

> "We sort of started a project, a couple of years ago about imprints. So in the UK if you distribute companion materials on paper, you have to say who is from whereas online that's not the case. And during the referendum and.. election.. where certain information comes from online, it wasn't attributed to campaigners in the same way that it would be offline. And so we tried to get some parliamentary support to change those rules." - P9

6.5.2.3 Organizing literacy campaigns

The advocators actively organize literacy campaigns educating people on how to critically examine information that they find on online platforms. For example, P5 shared how they partner with community radio stations, design MOOC [27] courses, and frequently share tip sheets explaining to people what fake news is and how they can identify it.

> "We try to also work with community radio stations .. and talk through what like what fake news is,.. talking about information literacy and information disorder and how it manifests" - P5

6.5.2.4 Spearheading initiatives to improve accessibility and quality of data and information

Better decisions are made when better data is available. Quality data and information is essential because it acts as a source to verify facts in the *short-term claims centric* fact-checking. The advocators are actively working to improve the code of practices in releasing data, for example, updating it from the paper to the digital age.

> "Improving the code of practice for official statistics and updating it to the internet age. And we've done lots of individual pieces of work trying to improve specific statistical releases because obviously, they form the basis of a lot of what we do.. we try and get ..the Office for Statistics regulation to be a bit bolder in how they treat misuse of statistics officially." - P9

The interviews also revealed that data in most of the African countries are either old or is not accessible. The advocators there are engaging with government agencies, making them aware of the problem and stressing the importance of having data publicly accessible.

> "Before the census in 2019, the last census had been done in 2009. So while there are estimates available on the census data, we would use those estimates. But then the data is just not accurate when it's estimated.. [So we] talk to the people at the National Statistics Office and say we would like this data. [We] talk to people at the ministry, and trade unions telling them that this is how it would be better if you track unemployment. " - P12

6.5.2.5 Building collaborations and coalitions

Several advocators stressed the need for all fact-checking organizations to work together, collaborate, and share resources. Such collaborations would be helpful in understanding the common challenges and needs of the fact-checkers. The coalition would also better position the organizations while making certain demands from the internet companies. IFCN has played a huge part in forming such a collaboration and several advocators are actively working to expand this network.

> "There's not a lot of us working together. So that's what I'm working on in collaboration with IFCN.. how different organizations and sectors can work better together to complement each other and collaborate and kind of information sharing and resources.., [understand] common challenges. For example, when we work with internet companies, there are certain things we might all want to ask them for which, at the moment, [only] some of us asking." - P9

In addition to a coalition among themselves, a few fact-checking organizations are also actively partnering with companies like Google and Facebook to fight fake news on their platform [102] "*because they have a huge impact on the way people experience misinformation* (P23)". P9 and P12 informed us how Google is working with their organizations to make their fact-checks more visible by ranking them higher in Google searches. One of the study participants (P18) was in fact instrumental in starting the initiative.

> "When we want what our fact checks to rank higher or to be more visible, so there is a back end tool claim review that we use that is integrated into WordPress.. And so the search engine looks at whatever you post as a fact check and then it ranks it higher in the matrix." - P12

6.6 Needs and Challenges of Stakeholder Groups

In this section, I answer RQ2 by presenting the challenges faced by various stakeholder groups categorized by emerging themes.

6.6.1 Skepticism Towards AI and Automation

Fact-checkers expressed skepticism and distrust towards artificial intelligence (AI) and automation of the fact-checking process. This finding resonates with prior work that

revealed that the black box nature of artificial intelligence techniques and machine learning algorithms make their inner workings unintelligible to humans thereby decreasing users' trust in their outputs [302, 348]. P4 divulged that they are skeptical of Facebook's AI-based tool that aggregates potentially misinformative content for fact-checkers to verify and rate on its platform because they think that "*algorithms hide a lot of stuff*". They believe that fact-checkers or independent organizations should be responsible for aggregating content to fact-check on the social media platforms rather than "*[companies] that are running the platforms*."

Fact-checkers understand that machine learning models work when there are lots of similar data for training and pattern recognition. Thus, P18 doubts AI's capability to detect falsehood in politicians' statements which can be very diverse in language and topics. They also doubt AI's capability to differentiate between a true and a false statement especially when there are only subtle differences between the two. P19 does not believe that fact-checking could be automated since its a complicated process that requires several humans to discuss and make decisions.

> I think [AI based tools are] going to be less useful for most.. politicians, because the problem is people don't repeat stuff the same way and the addition of a word or two can make a huge difference.. I don't think a computer is ever going to be able to figure that out. - P18

> "[Fact-checking is] such a complex process, for example, extracting a claim,- what's the underlying meaning of a certain claim, how to understand it. Even [for humans] it's a process of discussing and then deciding. So I think it still will be humans work in a way. " - P19

Despite the skepticism towards AI, I found that some fact-checking organizations have indeed adopted AI-based tools and a few others showed a willingness to adopt such tools. However, such tools are only acceptable for low-stake scenarios of monitoring content on social media as compared to high-stake scenarios of assigning a veracity label to the content. This observation is in line with recent work that found that the use of AI is more acceptable in low stake compared to high stake decision-making processes [56, 346].

> "I will not trust any algorithm or any AI to flag something as right or wrong, at least not at this stage.. I prefer AI only to give us a curated list, flag us that this is something that we should look in. Make a tool that picks up the signals that [indicate the content is] misleading and makes a curated list. " - P22

> "If you're serious with fact-checking, you cannot replace it by automatization or things like that. But it could be helpful in [social-media] monitoring. " - P19

To increase fact-checkers acceptance of AI-based tools, P16 and P12 stressed how it's essential to have humans in the fact-checking loop. P12 further said that they would only trust the AI output if the tool is able to explain how it arrived at a particular conclusion. Past studies have also developed human-in the loop AI systems [363, 430] and tried to make them more explainable in order to foster trust in them [35, 111, 258, 278, 348].

> "I think at a certain point, AI stops and human being needs to come in to verify.. I don't think that AI can really do exactly the same what we can do. Not yet at least." - P16

> " The manual process would still have [to be there]. I would be willing to use it [automated tool] to see how it arrives at that conclusion. So if you say this is misinformation, and these are the sources of data that we're using to make that. And we check and find that the algorithm.. is not using them out of context. So at that point, we would be in a position to say let's check it. " - P12

6.6.2 Need For Tools and Limiting Social Media Affordances

I found stakeholder groups divulging several challenges related to the tools they used and the affordances provided by social media platforms. In the process, I also discovered a few needs with respect to tools and systems. I present them next.

6.6.2.1 Monitoring Social Media Platforms is Manual, Time Consuming, and Difficult.

Fact checkers complained about the information overload problem. With the emergence of a variety of social media platforms, the amount of online information is increasing exponentially. However, for fact-checkers, searching for misleading content mostly remains a manual task. In addition, the generation of search queries that could lead to potentially dubious content is still based on the hit and trial method. As discussed in Section 6.4.3, only Facebook has provided a tool to its partner fact-checking organizations that can aggregate potential misinformation. While a number of social media monitoring tools have been developed to identify and aggregate misinformation, particularly on Twitter (such as [48, 88, 204, 327, 356, 389], etc.), other social

media platforms (e.g. YouTube, Google, Yandex, etc.) where misinformation is equally prolific also need attention.

> "The search is not easy to be honest.. it takes a lot of time to actually find the content because we only have a manual method to do that, hit n trial method to do that.. I have even scrolled to the point where YouTube shows me no more results. So that's how manual it gets." - P7

> "If there's some.. tool that helps filter..different types of misinformation will help because now you have information overload and you don't know what to choose and what not to choose." - P3

6.6.2.2 Limitations of platform affordances.

Many fact-checkers complained about how online social media platforms' affordances become a hindrance for searching and filtering content. For example, the inability to search for posts and comments on Facebook, a lack of search trends feature on platforms (with the exception of Twitter), the unavailability of fine-grained search filters, the inability to download content on platforms like Instagram, and the inability to search for same videos that were uploaded with different keywords (title, description, etc.) on YouTube are some of the limitations of platform affordances.

> "Facebook is actually tricky to be honest, because there's no one way to search content. you can only search people..pages and groups, etc.. but I need user generated content." - P7

> "Youtube search engine is not very good. It will just show you only videos which are popular which has no use. Misleading videos won't have too much views, but they have a lot of uploads. What happens is that since a lot of people.. upload videos using different caption, different keywords. So.. [there might be] 10 versions of the same thing. " - P8

6.6.2.3 Systems and tools needed to detect misleading claims on private message platforms.

Privacy settings on social media groups (e.g. Facebook) and end-to-end encryption in messaging platforms (e.g. WhatsApp) act as a hindrance in accessing misleading content circulating on these platforms. The fact-checkers informed that they need tools that allow them to access and flag content on these platforms. Recent research work

has focused on building crowd-sourced WhatsApp tip lines for discovering content to fact-check [230, 273]. However, being able to track or report private messages via tools necessitates serious ethical considerations. Private messaging platforms give users a false sense of security, thereby making them share sensitive information (e.g. clinical records of patients [267]), without anonymizing it [182] and hence, expose users to privacy risks [60].

6.6.2.4 Overload of tools.

During the interviews, fact-checkers showed and talked about several tools that they use during fact-checking. One of the participants P8 showed us around 15 tools. There is a tool overload problem. P3 suggested building a single tool that could provide all functionalities that fact-checkers need. However, since fact-checking is a complex task involving multiple steps, it is difficult for a single tool to cater to a divergent set of requirements and functionalities [389]. Thus, building a suite of purpose-built tools that cater to the specific needs of fact-checkers in the various steps involved in fact-checking would be more useful for the fact-checkers.

> "[If one could] put all these tools in one, .. So I don't have to look for different tools when I'm analyzing a video or an audio. I can do it in one place instead. Like.. I can use 10 tools to analyze the video. But if we can have one..[tool that] gives you the information you need." - P3

6.6.2.5 Need for specific tools

Despite the problem with tools overload, existing tools lack task-specific functionalities. For instance, a lot of steps involved in video fact-checking are manual. While tools like Invid and reverse image search are available to verify a video, they are only useful to check if the video is digitally altered or used in a different context than the original. For all other videos (e.g. videos with conspiracy theories), the claim extraction and claim verification process is manual. For lengthier videos, this process could get even more tedious.

> "There is no tool to debunk [conspiracy theory videos]. For example, I cannot do a reverse image search, and I cannot divide the video into keyframes, because it is a narrative that is false.. We have to really go line by line and see what the person is saying, and then all we can do is search quotes from different organizations to [verify the claim] ." - P7

Fact-checkers believe that "*efforts against misinformation advance much more and much quicker in English" (P23)* than in other regional languages. They need tools to transcribe videos in local regional languages.

> "The problem with transcribing videos online or using any other software is languages because India has so many languages and we tend to get a video and information in every possible language. So it becomes difficult to have one dedicated tool to transcribe all our videos." - P14

Editors revealed that editorial work is mostly manual.

P1 spoke about how editing work is "*still very human*" and human resource management tools like Trello could be further strengthened. For example, currently, "*there are a lot of issues of accountability*" with the Trello board, it lacks control features because of which "*anybody can move a card anywhere*."

6.6.2.6 Getting organic engagement for fact-checks without advertisements

P17 revealed that while on some platforms (e.g. Twitter) it is easier to *"get organic reach and engagement"* by adopting appropriate strategies, it is *"a big challenge to get [same].. attention without advertising"* on other platforms (e.g. Facebook).

6.6.3 Issues around policy and information infrastructure

Stakeholder groups discussed several challenges surrounding information availability and quality. They understand that quality data is essential not only for developing AI-based automated tools but also for investigating claims.

6.6.3.1 Need to improve information quality before automation

AI models are as good as the data available. If the information against which the claims are to be verified is missing or of low quality, the models would never work.

> "Automation is tricky because.. in a place like East Africa ..information is not readily available. You cannot say that I'll go to this site and get this information so that when these numbers are presented it can easily be automated." - P12

P9 raised an interesting point along the same lines. They explained how the success of automated fact-checking is dependent on the accessibility and format of statistics and information available. The data has to be in the same format for the machine to be

able to understand it. Their organization is working with institutes around the world to improve statistics globally.

> "[We want] statistics being published in a kind of open accessible and consistent format. So, for example, like any symbols that are used to show a caveat about data needs to be the same so that our machine can understand them each time.. My colleague.. is working with the open data Institute in UK and also globally.. to [determine] whether there are ways of improving statistics so that automated fact-checking can work." - P9

P22 elicited how data in their country is not available in a user-friendly or machine-usable format.

> "The data, you will find that it is in a PDF... file or a photograph on the website..then how do we use it?...You need the data ..[to be] put into a excel sheet so you can clean it. Data should be in a format which is user friendly so it can be used, [it should be] machine learning friendly so that the machine can pick up and use the data. " - P22

The aforementioned concerns elicited by fact-checkers are in line with the prior research that has also listed the absence of structured and quality data as well as lack of adherence to data standards, as few of the major challenges faced in the field of big data [42, 115]. Lack of harmonization between data sets makes data integration a complicated and time-consuming task [8, 9]. Data integration is necessary in various scenarios related to fact-checking [251], for example, querying different databases originating from different sources to determine the veracity of content or determining whether a piece of content has already been fact-checked by searching in various fact-checking databases [86, 251]. Thus, scholars have stressed on the need to adhere to common data standards to help facilitate data integration and reuse [67]. There have been a number of cross-country initiatives to set common standards for data in various domains. For example, in 2017, European Medicines Agency held a meeting to discuss the opportunities and challenges in applying a common data model to healthcare data across the countries in Europe to support regulatory decision-making [10, 11]. Furthermore, several advocators (Section 6.5.2) have also been spearheading initiatives to improve the data quality in their respective countries.

6.6.3.2 Lack of information sources.

The interviews with fact-checkers in the Global South revealed that the information needed to investigate claims either does not exist or is not updated periodically.

> "There's a lot that we don't cover because of lack of sources." - P2

> "In Kenya,.. demographic Health Survey, which .. shows the health situation in the country, the last one was done in 2014, this is 2020. It's just too old, and we can't use it." - P12

6.6.3.3 Difficulty in getting information for research from civic organizations.

In most cases, fact-checkers need multiple sources to debunk a claim. To obtain these resources they rely on public data sets and information from government and civic organizations. Six fact-checkers working in African countries and the Balkan regions informed us how information needed for research is not publicly available and getting it from officials is a long and difficult process.

> "It's so hard to get information from the government.. because everybody wants to protect themselves. They don't want to give you the information that you actually need." - P3

6.6.3.4 Algorithmic bias against local content

P5 complained about algorithmic bias in terms of how content indexed in search engines is skewed towards the Global North region making it very difficult to search for local content in regional languages of the Global South.

> "The way the algorithms works was very I'd say Euro-centric or like North America [centric].. Trying to find tools that would enable me to find.. stuff that's not in English, like things in local languages was quite challenging.. it would take a while for some of the stuff from from this side to be indexed on.. Google searches and other platforms .. So, there's sort of algorithmic bias when it comes to.. find stuff like that." - P5

6.6.4 Emotional cost of fact-checking

In addition to the manual labor involved in fact-checking, fact-checkers also face significant emotional toll and stress in their job. They are often victims of online

threats and abuse from users and conspiracy theorists whose posts they were tasked to debunk. Manually scanning through misinformative content about certain topics, such as riots, and conspiracy theories, also has adverse effects on their mental health.

> "It becomes kind of very stressful job. Seeing all violence and getting into each and every detail, kind of takes a toll on your mental health. And then again you have to listen to abuses. And the situation is worse when you like put this content online and then people attack you" - P8

> "We are exposed to threats, these conspiracy theorists are very aggressive and I worked only for like a couple weeks when I saw a CrowdTangle post with my own photo saying, this is your censor and it was very unpleasant feeling to find that." - P16

While previous work has studied emotional labor and psychological symptomatology in content moderation work [54, 123, 231, 318, 329, 369], no study has investigated the emotional cost of fact-checking work. Studying human costs underlying the fact-checking process and determining wellness interventions to psychological effects of fact-checking are other fruitful avenues for future research.

In this section, I discuss how this study renders visibility to the human and technological infrastructures supporting the fact-checking work and the collaborative efforts involved in the process. I also discuss the needs of the stakeholder groups and the implications of my findings on future research directions in fact-checking.

6.6.5 Rendering visibility to the human infrastructure of fact-checking

To date, the primary objective of fact-checking is considered as debunking misleading claims. Based on this objective, prior work suggests that fact-checking can only influence three constituencies— people, journalists, and political operatives [49]. By identifying and examining the human infrastructure—the stakeholder groups that need to be brought into alignment to accomplish fact-checking, this work provides a means to think about fact-checking as a multidimensional initiative which then assists in understanding 1) the invisible aspects of the process, and 2) the other ways the fact-checking process can have an influence. This work shows that fact-checking is supported by several processes, such as editorial work, social media engagement work, in-depth research and data analysis, as well as advocacy and policy work that might not be visible to the external world. Through the study of these processes,

I establish how fact-checking has evolved to include both *short-term claims centric* and *long-term advocacy centric* fact-checking. I make visible the efforts that fact-checking advocators are putting in to improve the availability, accessibility, and quality of data and statistics by aligning the focus and interests of governments and internet companies with fact-checking organizations. Through these efforts, the fact-checking organizations are not only improving the information landscape of their country but in turn are also improving the quality of (*short-term claims centric*) fact-checking itself. Rendering visibility to the work of the human infrastructure and the invisible processes of fact-checking has helped in uncovering the needs—both social and technical—of the entire fact-checking ecosystem consisting of all the stakeholder groups. The knowledge of the needs of the ecosystem could enable the design and development of tools and policies to support various aspects of the fact-checking process.

6.6.6 Collaborative efforts in the fact-checking process

This study highlights fact-checking as a distributed problem where collaboration takes place at multiple stages among people with different skill sets within and outside the fact-checking team/organization. First, collaboration occurs among the stakeholder groups: editors, fact-checkers, social media managers, researchers & investigators, and advocators (refer Figure 6.1 for an overview). Second, collaboration extends to the outside world with experts such as doctors, oncologists, academics, etc. whose expertise is needed to investigate dubious claims. Third, fact-checkers collaborate with civic and government organizations during the investigative stage to access data and statistics related to the claims under investigation. Fourth, in parallel, advocates reach out to policymakers to influence policy by highlighting the challenges faced by fact-checkers. Fifth, several internet companies, like Facebook, collaborate with fact-checking organizations to fact-check and debunk misleading claims on their platform. Finally, collaboration occurs between social media users and the stakeholder groups at two stages: 1) at the content monitoring stage where users report dubious claims that they encountered online directly to fact-checkers via tip lines, and 2) when users engage with fact-checks disseminated by social media managers.

Rendering visibility to the collaborative efforts in the fact-checking process has multiple benefits. I discuss a few.

6.6.6.1 Increase in efforts to foster collaborations

Making visible the collaborative efforts in the fact-checking ecosystem can lead to efforts and policies that foster these collaborations. There has been growing research on how to make users engage with the published fact-checks [50, 96, 132, 151, 432]. For example, fact-checking organization Pesacheck created a Twitter bot named *debunk bot* that detects tweets containing URL(s) to misinformative content and replies to them with the link to the fact-check that their organization has published [20]. Similar investigations and efforts can be put to support other collaborations, such as between experts and fact-checkers. There has been only a handful of recent efforts in this direction, for example, Meedan's Digital Health Lab[7] and Facebook's Journalism project[8] support fact-checkers in debunking health-related misinformation by connecting them with health experts and providing them with resources on the health topics that they are covering. Imagine a private social media platform consisting of separate communities (like subreddits) of fact-checkers, and experts from different fields (from doctors, journalists, meteorologists, to university librarians and professors) to facilitate easy and targeted communication and information sharing. Fact-checkers can post questions in relevant communities, seek quotes from experts, and get suggestions for online and offline resources to support their investigations. Fact-checkers can also share with each other their concerns or information about new tools that they discovered. Such a platform could facilitate fact-checking organizations in addressing online misinformation effectively and in a timely manner.

6.6.6.2 Revealing the value of fact-checking work to internet companies

In recent times several platforms such as Google search, YouTube, and Google images have started actively using fact-checks produced by fact-checking organizations with their search results to help determine their validity and truthfulness [19, 23, 29]. The study also revealed how fact-checking organizations are collaborating with internet companies and allowing them to index their fact-checks. This finding contributes towards the HCI scholars' call of making people aware of "the value their data brings to intelligent technologies" [24, 407]. This work highlights the value that fact-checking is bringing to social media companies that regularly use fact-checks to regulate the

[7] https://meedan.com/digital-health-lab

[8] https://www.facebook.com/journalismproject/facebook-partners-with-meedan-digital-health-lab-to-help-fact-checkers-debunk-health-misinformation?locale=pa_IN

content on their platforms and provide reliable information to the users. Previous scholarly works have raised questions on whether volunteer-created content such as Wikipedia articles should receive more economic benefits from the internet companies [405, 407]. Along similar lines, I want to raise the question of whether fact-checkers and fact-checking organizations should also receive more economic benefits for their work.

6.6.6.3 Revealing power dynamics in collaborations

This study also sheds light on the power dynamics of collaborations happening in the fact-checking ecosystem. For example, this work shows how editors have the power over fact-checkers in determining what kinds of misinformative claims should be prioritized. Additionally, in certain situations, fact-checkers depend on civic and government organizations to get data for investigation. Tools released by social media companies, such as Facebook Queue (refer Section 6.4.3.1) also dictate what kind of claims fact-checkers investigate. These companies also have the power to remove the fact-checks from their platform. For example, Facebook removed a fact-check on abortion after receiving complaints from Republican senators [21]. Further investigation of the potential consequences of these power dynamics in the fact-checking process is a fruitful avenue of future research.

6.6.7 Implications for future research on fact-checking

Taking a multi-stakeholder perspective on the fact-checking process helped us learn the needs of stakeholder groups as well as uncover the challenges that go beyond the technical aspects of fact-checking. I use the findings to discuss and propose various directions that future research on fact-checking can take. In Section 6.6.7.1, I start by discussing the needs of various stakeholder groups and propose solutions for the same. In Section 6.6.7.2, I discuss the values that the stakeholder groups desire in the tools and systems built for them. Next, in Section 6.6.7.3, I discuss how focusing on technical solutions alone is not enough and how the existing automated approaches fail to work in real life since they ignore the social aspect of fact-checking. I also reflect on the social and civic challenges faced by fact-checking organizations by discussing the role of information infrastructure in fact-checking. Finally, in Section 6.6.7.4, I end by stressing how the current research on misinformation has not focused on the Global

South countries and how there is a dearth of fact-checking tools built for regional languages of the Global South.

6.6.7.1 Technical needs of fact-checking

The study reveals that monitoring social media platforms is the most challenging aspect of the job of a fact-checker. A combination of over-reliance on third-party tools to discover potentially dubious content, limiting platform affordances, and a manual way of going through each search result to determine if it's potentially misleading makes the process extremely tedious. Access to better search filters on social media platforms, and a community-based approach to reporting misinformation where users of social media platforms are able to report problematic content are some useful ways to assist fact-checkers in finding misleading claims.

Fact-checkers also agreed that it's impossible for them to have access to all corners of the web. For example, they do not have access to content on private messaging platforms like WhatsApp that have become a popular haven for groups interested in sharing misinformation [47]. Given fact-checkers' skepticism towards AI and automation, user reporting via tip lines appears to be a feasible solution to access content on such platforms. Research on how to motivate people to report problematic content is a fruitful avenue for future research. I also found that fact-checkers end up viewing long videos to extract misleading claims and reading through the lengthy comments sections to get clues that would help them investigate the claims. A system that could utilize comments to highlight all the misleading claims present in the video, leaving the final decision of selecting what claims to verify to fact-checkers could be useful for the fact-checking community. Furthermore, a tool that highlights credibility indicators (such as comments containing useful information for investigating the claims) would significantly reduce the manual effort put in by fact-checkers.

The interviews revealed that when it comes to assigning a verdict to a claim there is no single standard labeling system shared across fact-checking organizations. Collective agreement on the labels could allow for greater information sharing and interoperability. Future efforts can focus on developing structured ontologies for representing credibility assessment metrics that would lead to the development of common labels and benchmarks for assigning veracity labels to the claims.

This work also sheds light on the needs of stakeholder groups other than fact-checkers. There is a lack of editorial and process management tools that could be used by news desk and copy editors. Social media managers need effective strategies to

increase users' engagement with fact-checks. In order for fact-checks to really have an impact, they must be "seen and attended to by audiences" [50]. To accomplish this, it is essential to understand who shares fact-checks on social media platforms and what modality or visual storytelling technique is more suited for which platform. Tools are also needed to convert fact-checks to multiple languages to increase their reach. Platforms can also help in making the fact-checks more visible and accessible. Google's efforts (e.g. claim review) to prioritize the ranks of fact-checks in searches is one such effort. Technology critics have called structured journalism, where fact-checks are produced in a machine readable form, as the future of fact-checking [41]. Recent work has tried to automate extraction of structured information—claim, claimant, and verdict from the fact-check, to allow search engines to display it in the search results [215]. More such efforts towards structured journalism are needed to integrate fact-checks with online content.

6.6.7.2 Values desired in fact-checking tools and systems

The study participants expressed skepticism about automation and AI technology because of its black-box nature. At the same time, they also showed a willingness to adopt automated solutions for low stake tasks. Fact-checkers do not want systems that decide the veracity of information, rather they want tools that could help them with their day-to-day tasks such as monitoring online platforms, checking whether a video is digitally altered, transcribing videos in regional languages, etc. Algorithm explainability combined with tools that have humans in the loop emerged as key values that fact-checkers desire in the systems built for them. Familiarity with the inner workings of algorithms along with tools that use both human and machine capabilities for problem-solving can help increase fact-checkers' trust in the automated systems. Recent times have witnessed a burgeoning interest in the field of human-centered XAI where researchers draw from formal HCI theories to design explanations on how machines reach a particular decision [36, 40, 55]. Scholars are also studying how to design human and machine configurations to operationalize human-in-the-loop systems [180, 430]. Understanding specific needs for explainability with respect to fact-checkers, operationalizing those needs at the conceptual and methodological levels in tools developed for fact-checking, designing systems that could take fact-checker's feedback and use that to modify the algorithm used by the system are few useful directions for future research.

6.6.7.3 Going beyond the technical: need for socio-technical solutions

Would technology-mediated solutions alone lead to improvement in the quality of fact-checking? While automating the entire fact-checking process and developing new tools to increase efficiency and scalability seems promising, it is not a panacea to all the problems faced by the stakeholder groups. A holistic change could only be achieved via systematic changes in the civic, political, and informational contexts. Through this study, I found how accessing data from the civic and government bodies is a difficult task, especially if it portrays the government in a less-than-favorable light. While fact-checking is now a global endeavor, in some countries information is either not publicly available or essential information (e.g. census data, health surveys, etc.) is outdated because of a lack of periodic collection. The interviews revealed how several claims are left unchecked because of a lack of sources. The availability of high-quality up-to-date information is essential for fact-checking—manual or automatic. Additionally, good quality data is a precursor to having good machine learning models [355] that would be needed to automate the investigative step in the fact-checking pipeline where publicly available authoritative statistics are used to determine the veracity of claims. Thus, as part of *long-term advocacy-centric fact-checking*, advocators within the fact-checking organizations have been actively pushing for policy changes to improve the availability and quality of data and statistics. For example, fact-checking organization Full Fact gave oral evidence to the House of Commons Public Administration and Constitutional Affairs Committee on issues of coherent and accessible health statistics in the United Kingdom [296]. There is a growing interest in the HCI community to engage with policymakers as a way to inform policy that could benefit society [248, 365]. Future work in fact-checking could focus on understanding the opportunities and difficulties that advocators face while engaging with civic organizations and determining strategies that advocators can adopt to shape policies surrounding statistics and public data in their respective countries.

6.6.7.4 Fact-checking in the Global South

Recent work in the CSCW community stressed on the fact that academic research, to date, has primarily focused on misinformation in Western countries, while not addressing the phenomenon in the Global South [190]. This study reiterates the lack of knowledge and context surrounding the misinformation landscape in the Global South. Local regional languages are under-resourced both by online platforms and

search engines making it difficult for fact-checkers to gain access to local context online and for regional speakers to gain access to reliable information [126]. The current design-based approaches to fact-checking do not take into account the lack of fact-checking resources in regional languages and thus, fail to account for the unevenness of the viability of fact-checking across the globe. Advocators recognize how access to local community-specific knowledge and culture is essential to understand the characteristics of misinformation and why it spreads in a particular region [144]. Thus, they are calling attention to the need to improve the information infrastructure and research in the Global South. This work also acts as a call to action for researchers to study the misinformation landscape in the Global South region.

6.7 Conclusions and Limitations

This work sheds light on how fact-checking is practiced in the real world by presenting the infrastructures—both human and technological—supporting the fact-checking work. I interviewed 26 participants belonging to six primary stakeholder groups involved in the fact-checking process namely editors, external fact-checkers, in-house fact-checkers, investigators and researchers, social media managers, and advocators. By studying the various tasks performed by these stakeholder groups, I identified the role of tools, technology, and policy in their work. Finally, I also identified key challenges faced by the stakeholder groups along with opportunities of advancing current tools, policies, and technology for fact-checkers.

This work is not without limitations. The majority of the organizations that I interviewed are either IFCN signatories or work closely with IFCN signatories. Thus, the fact-checking model and tools used by the stakeholders presented in this study might not apply to every fact-checking organization. The fact-checking process could also be subjected to various degrees of local and regional variabilities that this work does not capture. I also acknowledge that not all stakeholder roles are present in every fact-checking organization that I interviewed. For example, not every organization is doing long-term investigative or advocacy work. I leave the examination of the factors influencing the fact-checking work in different regions and the variability in fact-checking work across regions to future work.

CHAPTER 7

DEFENDING AGAINST ONLINE MISINFORMATION VIA SYSTEM DESIGN

In Chapter 6, I interviewed fact-checkers to understand the process of fact-checking, and the technical, informational, and social needs of the fact-checking organizations. Using the insights from that work, in the last thread of my research, I design and build a system to combat online misinformation. More specifically, I present *YouCred*—a fact-checking system that I designed with and for fact-checkers to help them with their fact-checking workflow on the YouTube platform.

7.1 Motivation

With the growing scale and widespread dissemination of online misinformation, monitoring social media platforms has become an ongoing challenge. While there has been an increase in efforts to develop fact-checking systems, especially for platforms like Twitter and Facebook, these systems have struggled to make a substantial impact on the practices and reporting methods of fact-checking organizations [317]. Moreover, there is a noticeable absence of such tools for video search platforms such as YouTube [317]. This is particularly noteworthy considering that YouTube is the second-largest search engine and the most popular video-sharing platform [3] and has been frequently labeled as a hub for conspiracy theories [45, 82, 416]. To address this issue, I partnered with Pesacheck [311], Africa's largest indigenous fact-checking organization,

to develop YouCred—a fact-checking system for the YouTube platform.

To ensure that YouCred caters to the actual day-to-day needs of fact-checking organizations and assists them in their fact-checking workflow on the YouTube platform, I first conducted a formative interview study. The objective of the study was to gain insights into the methods employed by fact-checkers for monitoring and fact-checking videos on YouTube, as well as to identify the challenges they encounter and their specific requirements. I interviewed seven participants performing various fact-checking roles (as identified in [317]) at Pesacheck. I also included two fact-checkers outside of Pesacheck to ensure the broader applicability of YouCred beyond a single organization. Through this study, in line with the findings of Chapter 6, I found that fact-checkers heavily rely on manual searches to find misleading content on YouTube. In addition, generating search queries that could lead to potentially dubious content is still based on guesswork, domain knowledge, and experience. As a result, fact-checkers spend countless hours crafting search queries and scanning YouTube in search of potentially misleading information. Furthermore, I also discovered a lack of video annotation tools to aid fact-checkers with credibility assessments on YouTube. Based on these insights, I designed the YouCred system to aid fact-checkers with misinformation discovery and credibility assessments on YouTube. The system is a result of a 2-year long collaboration with Pesacheck. I integrated the knowledge and feedback of Pesacheck's fact-checkers throughout all stages of system development, including requirement elicitation, feature engineering, system design, deployment, and testing. YouCred offers a misinformation discovery feature designed to generate search queries related to significant events and topics of interest to fact-checkers which are likely to yield misinformative results on YouTube. The system also provides an intuitive credibility assessment interface that simplifies video annotation, allowing fact-checkers to highlight misleading claims, add comments, and contribute other relevant information to investigate the veracity of the information presented in the videos.

To test the acceptance of YouCred, I deployed the system at Pesacheck for nine months and closely monitored its usage. I also conducted interviews with Pesacheck's team to gain deeper insights about YouCred's applicability. My evaluation revealed that YouCred was used until the very end of the deployment period. Fact-checkers found YouCred to be a valuable tool. It provided them with a wealth of information that would otherwise be challenging to obtain manually, enhancing their fact-checking capabilities on the YouTube platform.

Overall, YouCred represents a significant step towards combating online misinfor-

mation on video search engines. The collaborative efforts highlight the importance of an ongoing dialogue between fact-checking organizations and technology developers to ensure the relevance and effectiveness of technological solutions in the fight against online misinformation. I show that by incorporating stakeholders' insights in designing fact-checking systems, I can develop solutions that have a tangible impact on real-world fact-checking practices. My study makes the following contributions:

- YouCred serves as an example of how to design systems that bridge the gap between the needs of fact-checking organizations and the development of fact-checking systems. By integrating the knowledge and feedback of Pesacheck's fact-checkers throughout the entire development process, YouCred is designed to align with the values and fulfill the requirements of fact-checkers in a real-world context.
- YouCred introduces an innovative approach to assisting the fact-checking process by automatically generating search queries pertaining to important events and topics of interest. This feature provides structure to the current search query generation method which traditionally relied on guesswork and lacked a systematic approach.
- The YouCred system provides an intuitive interface for annotating YouTube videos, allowing fact-checkers to highlight misinformative claims, add comments, and other relevant information. This streamlined annotation process simplifies the identification and flagging of misleading content within videos, enhancing the overall fact-checking workflow.
- YouCred was deployed and monitored for a period of 9 months, allowing for extensive evaluation. The continuous usage of YouCred during this period indicates its practical value and usefulness for fact-checkers. Additionally, the study also underscores that designing and deploying systems is not enough, we need continuous maintenance and evolution of systems to effectively address the evolving needs of fact-checkers.

7.2 Formative study

To inform my work, I conducted a formative interview study to understand the current fact-checking practices to monitor YouTube and opportunities for improving those practices. In this section, I describe the details of the formative study including the participant details (Section 7.2.1), study procedure (Section 7.2.2), and findings

(Section 7.2.3). Next, I describe the design goals I identified to guide the development of YouCred (Section 7.2.4). Finally, I describe the collaborative and iterative design process adopted for the study (Section 7.2.5).

7.2.1 Participants and Procedures

I conducted in-depth, formative interviews with 9 participants. Among them, 7 participants were affiliated with Pesacheck (referred to as P1-P7) and held various roles within the organization. These roles included 4 external fact-checkers responsible for monitoring online platforms, investigating dubious claims, and creating fact-check reports, 2 news editors overseeing the fact-checking process, and an advocator engaged in long-term investigative research on persistently circulating misinformation. For a detailed understanding of these roles, please refer to [317]. To ensure the system's flexibility and adaptability to different organizations' needs, I also interviewed 2 additional external fact-checkers (referred to as P8-P9) from other fact-checking organizations (DPA [31], and First Check [30]). In order to preserve the participants' anonymity, I refrain from providing their demographic details and specific roles within the organization.

7.2.2 Interview protocol

I began the interviews by asking participants how they monitor YouTube and discover misleading videos on the platform. Then, I asked participants to share their screens and guide me through a fact-checking report debunking a false claim made in a YouTube video. I asked them to explain the entire process followed in the report, including how they found the video, identified the problematic claim, conducted investigative work, and wrote the fact-checking report. Next, I asked questions about the tools that fact-checkers employed in the process as well as any disadvantages associated with those tools. Finally, I asked about the challenges encountered throughout the fact-checking process on YouTube and discussed how the affordances of YouTube facilitated or hindered their work. Overall, the formative interview study provided valuable insights into the needs, values, and challenges experienced by the stakeholder groups. I conducted interviews on Zoom and they lasted between 60-90 minutes. The first and second authors went through the transcriptions of the recordings and coded them using thematic analysis. In the next section, I report the findings of the formative study.

7.2.3 Findings

All participants emphasized the challenge of finding and fact-checking misleading content on YouTube. One participant stated how "*they have seen a decline in the YouTube videos that [they] fact-check because of the hardships that [they] face with finding misinformation on the platform*" (P5). The interviews revealed that searching for misleading content on YouTube remains a manual and difficult task. The generation of search queries to find potentially misleading content is based on "*guesswork and trial-error method*" (P3). As one participant explained, "*To search about a particular topic/claim on YouTube, fact-checkers start with relevant terms related to certain claims or terms that best describe that person or their situation or history, and [they] keep trying different options until [they] find some results that they're looking for*" (P4). P3 further added that searching for generic queries like "*Obama will get both misinformation and non-misinformation content. So it's up to [the fact-checker] to through all search results to figure out which one is misinformation*". Going through the search results is also challenging because the top search results usually contain videos from mainstream news channels. As one participant elaborated, "*It's sort of hard for you to zero down to exactly what you're looking for... If you look for, say, Ebola, they'll give you these videos that are authentic, say, from the BBC, DW, Aljazeera, and CNN, and that is not potential misinformation*" (P5).

The modality of videos poses additional challenges, as fact-checkers may have to "*watch very long videos from start to end*" (P2). To save time, some fact-checkers prefer using transcripts and searching for specific keywords. As one participant explained, "*I go to a video that is 1 hour long and if you're looking for misinformation about COVID-19, I just search for COVID-19, and then I only go through the instance where the person in the video spoke about COVID-19*" (P3).

Interviews revealed that the YouTube platform also lacks certain affordances that are useful for fact-checking workflows, for example, "*the ability to search for a lot of things at once*" (P2), "*get the analytics for search results*" (P2), ability to download search results (P2, P3), and the capability to filter search results using a combination of search filters, for example, date-range and engagement metrics (P2, P3, P4, P5). The interviews further revealed that the process of annotating a video for veracity is unstructured and they rely on google docs to keep track of all information about the fact-check—"*When it comes to video annotation, it is very analog. What we do is take the actual video, download the transcript..figure out what part of the video we want and use google docs to annotate the videos*" (P1).

When probed about what kind of tool they envision, I repeatedly observed partici-

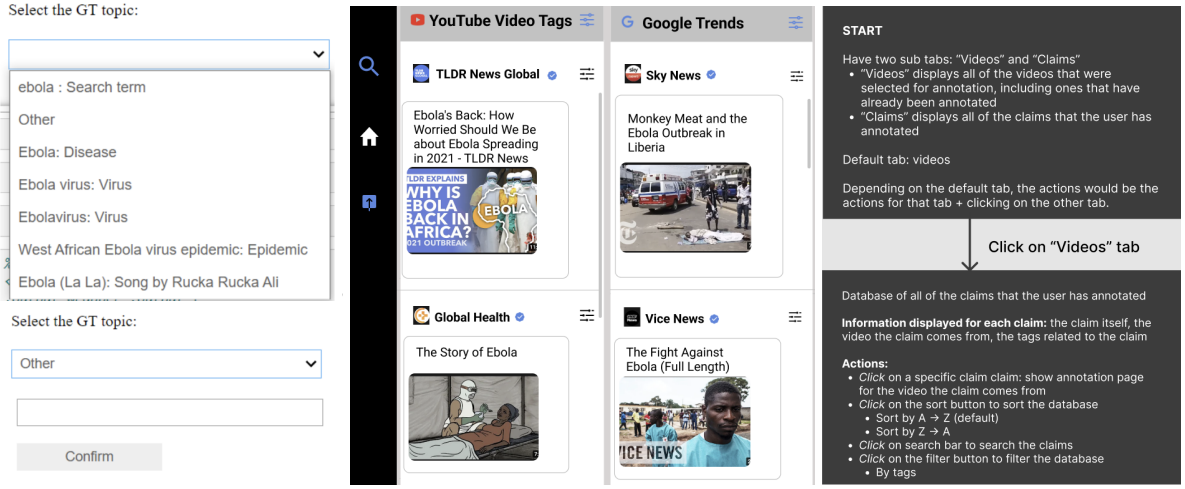
pants expressing to have control over the outcomes of the tool, for example, "*I want to be able to choose the keywords*", "*I need the ability to decide what videos I want to look into*", "*human being needs to come in to verify. I don't think that [any tool] can really do exactly the same what we can do, not yet at least*". All of these statements underscored the participants' strong desire for a sense of control and agency when it comes to the tools they would utilize. Participants also expressed a desire for the tool to support searching "*in a specific country, a specific region*" (P5) and the ability to search in "*regional languages of Africa, such as Amharic*" (P7). They also desired a unified system that incorporates most elements of their workflow, eliminating the need to switch tabs and providing a seamless experience (P4, P6).

7.2.4 Design goals

Drawing from my own research in Chapter 6 and the formative study, I identified five design objectives that inform the design of the YouCred system. These objectives directly address the needs and obstacles fact-checkers encounter when monitoring YouTube and conducting credibility assessments on YouTube videos.

- *Automated Search Query Generation:* YouCred aims to automate the generation of search queries, which is currently a tedious and manual task for fact-checkers. The system aims to eliminate guesswork by suggesting relevant keywords that could lead to misinformative search results.
- *Data Visualization and Insights:* YouCred empowers fact-checkers with intuitive and informative data visualizations that present crucial engagement patterns (such as likes, views, and comment counts) and publication dates in a clear and accessible manner. This feature enables fact-checkers to efficiently prioritize videos for analysis.
- *Multi-Term Tracking:* Fact-checkers expressed a desire to track multiple search terms and channels simultaneously. By providing this functionality, YouCred aims to facilitate efficient monitoring and analysis of content relevant to their topics of interest.
- *Adaptability to Different Languages and Contexts:* YouCred addresses the fact-checkers desire for multilingual content discovery and annotation capabilities by supporting search in various languages and in specific regions.
- *Agency and Human Involvement:* Fact-checkers value their agency in the fact-checking process and consider human involvement crucial for verification and contextual understanding. YouCred aims to strike a balance between automation

and human expertise, allowing fact-checkers to have control over verification capabilities while leveraging the benefits of automated tools.



(a) **(b)** **(c)**

Figure 7.1: (a) A snapshot of the UI widgets implemented in the Jupyter Notebook to demonstrate the search query generation methods, (b) Figure presenting the initial wireframe of the YouCred *view-results* page, developed in Figma, (c) Figure displaying an example snapshot of one of the initial workflow diagram created for YouCred

7.2.5 Design process

I adopted a collaborative and iterative design approach to create the YouCred system. From June to September 2021, I came up with search query generation methods based on prior literature and the findings of the formative study. These methods were implemented using interactive widgets in a Python Jupyter Notebook (e.g. Figure 7.1a). Iterative improvements were made based on feedback received from the fact-checking community at Pesacheck. The feedback was largely positive, and the fact-checking team expressed interest in supporting the design and evaluation of the system. From October to December 2021, the first author collaborated with two undergraduate students who had extensive experience in UX design and prototyping. Together, they created wireframes (e.g. Figure 7.1b) and workflow diagrams (e.g. Figure 7.1c) using Figma[1], a widely used prototyping tool. The development of YouCred commenced in January 2022, with features being added incrementally. Regular meetings were conducted with Pesacheck's team to showcase the built features and gather feedback.

[1]https://www.figma.com/

The team graciously shared their time, expertise, and ideas, demonstrating a deep and continuous engagement throughout the project. Between May 2021 and June 2023, I conducted approximately 38 meetings with the Pesacheck team. The meetings were attended by 1-21 members from Pesacheck, with the team lead of the fact-checking team being the sole attendee in four of the meetings. The UX designers and developers who were involved in building YouCred also joined every meeting. These meetings were conducted in English over Zoom and had a duration of approximately 1 to 2.5 hours. Additionally, constant communication was maintained through a Slack channel. In the next section, I describe the YouCred system in detail.

7.3 Overview of YouCred

YouCred is a fact-checking system specifically developed to support fact-checkers in combating online misinformation on the YouTube platform. This comprehensive system offers a wide range of functionalities that significantly enhance the efficiency and effectiveness of fact-checking activities. At its core, YouCred features a powerful misinformation discovery module that generates search queries that are of interest to fact-checkers and have a high probability of returning misinformative videos on YouTube. By eliminating the need for manual query formation, this automated search capability saves fact-checkers time and effort. Furthermore, YouCred incorporates a visualization component that provides detailed insights into date of publication of videos, as well as the engagement received by search results, including likes, dislikes, comments, and view counts. This visual representation empowers fact-checkers with valuable information to prioritize their investigation efforts more effectively. Additionally, YouCred offers a user-friendly credibility assessment interface, facilitating the annotation and analysis of YouTube videos. Fact-checkers can easily highlight misinformative claims, add comments, and provide additional information to delve deeper into the veracity of the presented information. This streamlined interface accelerates the fact-checking process and empowers fact-checkers to efficiently evaluate the credibility of videos. In the upcoming sections, I provide a detailed description of the design of the YouCred system and its functionalities, specifically focusing on how it facilitates misinformation discovery (Section 7.4) and credibility assessments (Section 7.5).

7.4 Misinformation discovery

The primary objective of the YouCred system is to help fact-checkers discover misinformation on the YouTube platform. The system accomplishes this objective by generating search queries related to important events and topics that need monitoring and are of interest to fact-checkers. Search queries are generated via four methods that include leveraging YouTube video tags, Google Trends search queries, YouTube's autocomplete suggestions, and analyzing frequently occurring words within the transcripts of misinformative videos. However, for each of these methods to work, fact-checkers are required to provide a small, curated list of seed misinformative (or potentially misinformative) videos pertaining to a particular topic as input. While only a minimum of one seed video is required for a topic, I recommended fact-checkers upload at least 3 videos for the topic to get better results based on my testing. I explain the ways fact-checkers can provide input to the system in Section 7.4.1 and expand on the search query generation methods in 7.4.2.

7.4.1 Inputting seed videos to YouCred

The fact-checkers have two options to input seed videos into the YouCred system: uploading a CSV file for each topic separately or utilizing the 'YouTube-CSV-Helper' Chrome extension. Initially, when the system was deployed in September, only the CSV method was available for inputting seed videos. However, after a month of testing and usage, fact-checkers provided feedback that creating input CSVs for each topic was a tedious task, hindering their frequent use of the system. To address this issue, I conducted two brainstorming sessions with the fact-checking team and developed the 'YouTube-CSV-Helper' (YCH) Chrome browser extension. This extension simplifies and automates the process of providing seed misinformative videos to the YouCred system. Once the seed videos for a specific topic are uploaded using either method, they are stored in a database and remain accessible for future use. Fact-checkers can conveniently browse through existing topics without the need for repetitive CSV creation and uploading. In situations where multiple fact-checkers upload CSVs or use the extension to submit videos about the same topic, I merge the videos and eliminate duplicates. The YouCred system also allows users to view, edit, delete, and download all uploaded topics and their corresponding seed videos as a CSV file.

Below, I provide a detailed description of both the CSV and extension methods of providing input to the YouCred system.

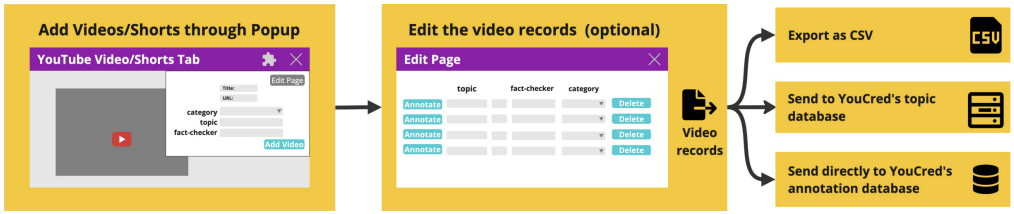


Figure 7.2: Figure illustrating the workflow of YouTube-CSV-Helper extension.

7.4.1.1 Manually uploading a CSV

YouCred offers fact-checkers the capability to upload a maximum of 30 CSVs simultaneously, with each CSV focusing on videos related to a specific topic. It is essential that each CSV includes standardized column headers, namely Video Title', Video Link', and Misinformation Category'. The Misinformation Category' aligns with the veracity labels currently employed by Pesacheck, encompassing categories such as altered, false headline, hoax, missing context, partly false, satire, false, and likely false. I also accommodate a 'No conclusion' category for situations where fact-checkers add potentially misinformative videos as seed videos that have not undergone a formal fact-checking process at the organization. To ensure consistent topic naming, fact-checkers are required to manually enter the topic for each CSV. To aid in this process and promote topic name consistency, I provide fact-checkers with a user-friendly drop-down list containing all existing topics. This prevents fact-checkers from inadvertently using different names for the same topic, such as "Kenya elections" and "Kenyan elections".

7.4.1.2 Via CSV helper extension

To simplify the process of gathering seed videos for YouCred, I developed the 'YouTube CSV Helper' browser extension. This extension offers fact-checkers a convenient way to store the details of potentially misinformative YouTube videos they come across on social media, user tip-lines, etc. When fact-checkers encounter such videos, they can simply click on the extension, triggering a popup that displays the video's title and URL. In this popup, fact-checkers can enter the misinformation category and the topic to which the video belongs. To ensure consistency, I provide a drop-down list that contains all existing topics. Additionally, fact-checkers have the option to add their name for proper attribution, although this field is optional.

The popup also includes an 'Edit' button that directs fact-checkers to a page listing all videos along with their associated topics, misinformation categories, and fact-checker names. Each video entry in the list is equipped with a 'Delete' button,

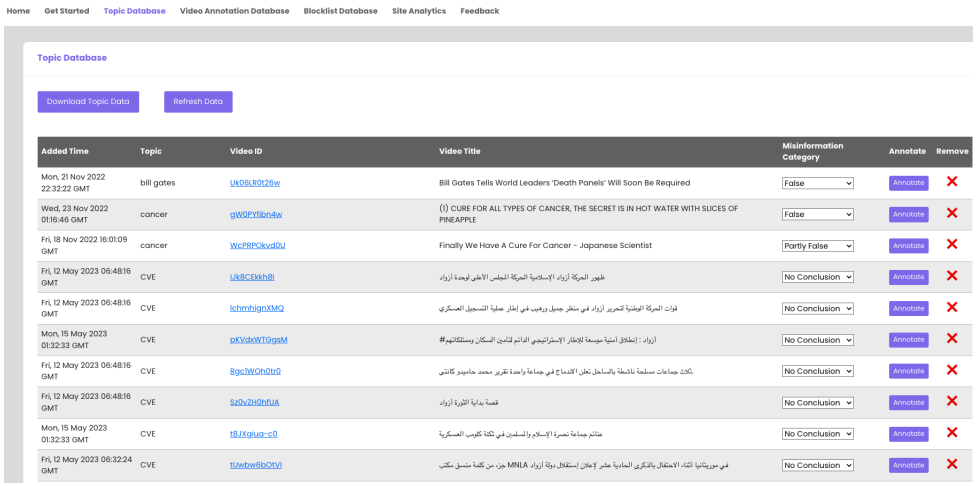


Figure 7.3: Snapshot of YouCred's topic database.

allowing fact-checkers to remove entries as needed. Furthermore, an 'Annotate' button is available for each video, enabling fact-checkers to send the video to the annotation database of the YouCred tool. This feature is particularly helpful if fact-checkers intend to fact-check the seed video(s) and publish a corresponding fact-checking report. The edit page further offers the option to filter videos by topic, providing fact-checkers with enhanced organization and accessibility. Moreover, fact-checkers have the flexibility to selectively send all or specific topics to the topic database and/or the annotation database of YouCred. Additionally, they have the capability to download topics and their corresponding videos as separate CSV files, facilitating easy access and storage of the data. These features collectively contribute to a more efficient and streamlined process for fact-checkers to provide input to the YouCred system. Figure 7.2 shows the workflow and features of the extension.

7.4.2 Formation of search queries

Fact-checkers can form search queries for topics for which they've uploaded seed videos. They can choose four methods of query generation including query generation using video tags of seed videos, frequently occurring words in YouTube videos, the Google Trends platform and YouTube's autocomplete suggestions. For each of the methods to work, fact-checkers have to input their YouTube API key in the tool. All four methods provide fact-checkers with complete agency in terms of what they want to search on YouTube. They are free to add/modify/remove search terms from the generated search queries or add their own search queries which can then be monitored

on YouCred. After finalizing search queries, fact-checkers can select optional region and language parameters[2]. The region parameter returns search results with videos that can be viewed in the specified countries (default value is 'Worldwide'). Language parameter returns search results that are most relevant to the specified language (default is 'All languages'). Both these parameters were requested by the Pesacheck team since they sometimes monitor videos specific to a region and language. Fact-checkers also have the ability to specify the date range to obtain videos that were published in that timeframe on YouTube. They have to enter the number of search results desired for the search query (minimum 1 and maximum 200[3]) and select one or more search filter(s) by which they want the search results to be sorted (uploaded date, video rating, relevance, video count, or number of views).

I added an additional feature to the search query generation methods, which allows for the exclusion of videos that have been marked as "blocked" by the fact-checkers. During the process of reviewing the search results (as explained in Section 7.4.3), fact-checkers have the capability to block videos that do not contain misinformation. Once this optional feature is selected, the blocked videos will not appear in the YouCred search results. By blocking non-misinformative videos, fact-checkers would have the option to focus their attention and resources on reviewing videos that are more likely to contain misinformation. This streamlines the fact-checking process and enables fact-checkers to allocate their time and efforts more effectively.

Fact-checkers informed me that YouTube's top search results predominantly feature videos from mainstream channels. They compiled a list of 46 credible mainstream news channels that have a lower likelihood of containing misinformation. Some examples include The Ethiopian Reporter, Nation Africa, KTN News, New York Times, and Politifact. The team requested to exclude videos from these channels when displaying the search results. Although not visible on the system, I internally remove videos from these channels by adding a minus operator [277] before each channel name in the generated search query. This modified query is then used to query YouTube and retrieve the search results. It's important to note that the capability to exclude blocked videos and remove videos from selected channels is available across all four query generation methods. I will now provide a detailed description of the four methods.

[2]https://developers.google.com/youtube/v3/docs/search/list

[3]The maximum number of search results was selected after consulting the Pesacheck's fact-checking team.

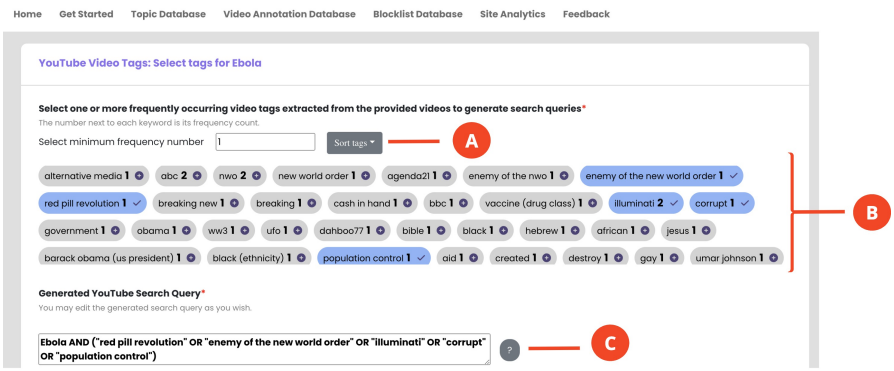


Figure 7.4: Figure illustrates YouCred's query generation method page, which utilizes the YouTube video tags method. The page displays a collection of video tags that can be sorted either by frequency or alphabetically (A). Each tag is accompanied by its frequency of occurrence. When a tag is selected, its corresponding bubble changes color to blue (B). Fact-checkers can choose multiple tags, and as they make their selections, the chosen tags are appended with the topic to form the search query. Importantly, the search query is editable, allowing fact-checkers the agency to modify it as needed (C).

7.4.2.1 YouTube video tags

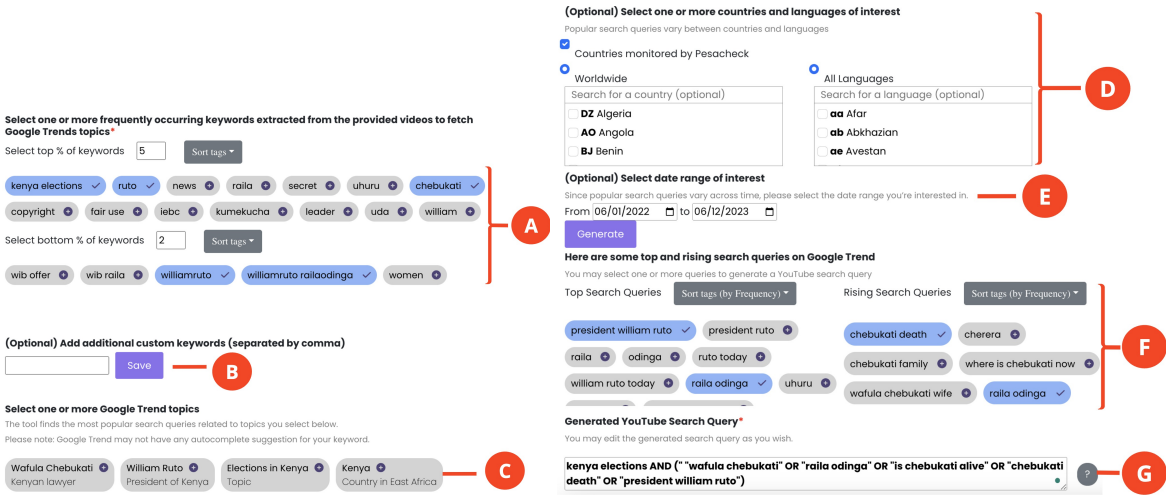
This method leverages YouTube video tags found in the seed videos uploaded by fact-checkers to YouCred. Video tags are chosen by the channel owners and are typically not visible to viewers. However, they can be accessed through the YouTube API[4]. These tags can be viewed as search words that content creators use to enhance the discoverability of their videos. They serve as labels that indicate the associated topics, themes, or relevant keywords of the video, helping users find the video more effectively on YouTube's search engine[5].

The tags extracted from the seed videos can also be employed as search queries to uncover more misinformative videos on YouTube. Previous research has demonstrated that video tags are highly informative input features for detecting misinformative videos [303]. Moreover, the practice of using tags as search queries, referred to as "misinfo-queries," has been utilized in the literature on misinformation audits, where tags were employed to retrieve misinformative results [222].

In YouCred, I empower fact-checkers to sort tags alphabetically or by frequency. As fact-checkers select the tags, I append them to the search topic using search query operators. In case the seed videos do not have any tags associated with them, fact-checkers would have the option to proceed to the next query generation method that

[4]https://developers.google.com/youtube/v3/docs/videos
[5]https://support.google.com/youtube/answer/146402?hl=en

they selected or return to YouCred's home page. Figure 7.4 shows the interface of YouCred query generation page via tag method. The seed videos have used tags such as 'enemy of the new world order' and 'red pill revolution', 'population control'. These selected tags are combined with the topic name 'ebola' to form a search query which can be used to find more misinformative videos related to Ebola on YouTube.



(a) (b)

Figure 7.5: Figure depicts YouCred's query generation page utilizing the Google Trends (GT) method. Fact-checkers begin by selecting keywords that serve as seed words for extracting GT topics (A). They also have the flexibility to add custom keywords (B). Next, fact-checkers choose the GT topics of interest (C), select the countries and languages (D) they want to focus on, and specify the desired date range (E). The system then extracts the GT search queries (F), which fact-checkers can review and select from. The search query generated is editable, allowing fact-checkers to modify it as needed (G).

7.4.2.2 Google Trends

Fact-checkers actively monitor online platforms by tracking current topics and real-time search trends of people [317]. They keep a vigilant eye on popular claims and assertions related to topics of interest [317]. To assist fact-checkers to monitor popular themes about a topic of interest, I leverage Google Trends (GT) platform. GT's search queries are a good indicator for understanding how people search for a topic on Google-owned platforms including YouTube. As a result, researchers have extensively used search queries obtained from GT to monitor misinformation and disinformation on online platforms [92, 205, 222, 223, 319, 338]. YouCred also utilizes GT's search

queries to help fact-checkers gain a better understanding of the public's interest and the prominence of certain claims on the YouTube platform. YouCred extracts and displays both the most popular and least popular search queries related to a topic in a specified region and time period on the YouTube platform. The system showcases the most popular search queries as they represent the commonly used terms by users. Additionally, the system presents the least popular terms, as these could potentially be exploited by conspiracy theorists to spread false information, a phenomenon known as data-voids [167].

Figure 7.5 shows how the process of obtaining search queries from GT is automated. As a user manually enters a seed word in the search bar of GT, the platform presents a dropdown list of existing GT topics containing that word. For example, entering seed word 'Ebola' results in GT suggesting GT topics such as 'Ebola virus', 'Ebola disease', 'West African Ebola virus epidemic', etc. To automate this step, I prompt fact-checkers to select a few relevant seed words. To assist in the selection process, I provide a list of the most and least occurring unigrams and bigrams found in the titles and descriptions of seed videos related to the topic. Once the seed words are selected, I curate all the GT topics suggested by the platform. Fact-checkers can then select the relevant GT topics from the list. Then, I ask fact-checkers to select the date range, country, and language of search trends. I utilize these parameters along with the GT topic to extract and present the search queries about the topic. All search queries selected by fact-checkers are appended by the OR (+) operator [6] to ensure that the search results obtained encompass the chosen queries.

7.4.2.3 YouTube transcript

YouCred utilizes the keywords occurring in the misinformative seed videos as potential search queries. A misinformative YouTube video could potentially have multiple false claims about a topic and the keywords associated with the claims could be potentially used as search queries for finding other misinformative videos related to the topic. To facilitate this, YouCred extracts transcripts of all the seed videos using the Python open-source package, *youtube-transcript-api*. The transcripts are subjected to standard text preprocessing steps of stop word removal and lemmatization. Then, I use TfIdf to extract the unigram and bigram features. TfIdf assigns higher weights to terms that are frequent within a specific transcript but relatively rare across the entire collection. This helps in distinguishing significant terms that are indicative of

[6] Search query search-term-1 + search-term-2 will return videos containing any of the search terms.

the content and themes of the seed videos. The usage of TfIdf features in the analysis of misinformative YouTube videos has been widely adopted and recognized in literature [140, 199, 222, 354]. The fact-checkers can choose from the top x% of the features (default value of x is 5% but could be modified). To form the search query, all features selected are appended together using the OR operator. Fact-checkers have complete agency to add/remove/edit and term of the query generated.

7.4.2.4 YouTube autocomplete

Autocomplete suggestion is a widely used functionality found in web search engines, designed to aid users in constructing their search queries. It is estimated that about 75% of search queries on Google are influenced by autocomplete suggestions [32]. Prior studies have shown that auto-complete suggestions could lead users to misinformation online [200, 205, 223]. Additionally, previous work has revealed that autocomplete suggestions in languages of the Global South, such as Amharic, Kiswahili, and Somali, could especially expose users to harmful content [100]. Given their impact, misinformation appearing in search results for these terms could impact a large number of people and thus, are of interest to fact-checkers. Therefore, YouCred enables fact-checkers to track autocomplete search suggestions for topics of interest.

In order to facilitate the selection of seed words for obtaining search query suggestions, I provide fact-checkers with the top and bottom x% of frequently appearing unigrams and bigrams. These are curated from the titles and descriptions of seed videos related to the topic at hand. By default, the value of x is set at 5, but fact-checkers have the freedom to modify this value as they see fit. Additionally, fact-checkers can include their own custom seed words for generating queries. YouCred extracts all search queries for the selected and/or entered keywords and prompts fact-checkers to select the ones that they want to monitor. The selected search queries are joined using OR search operator.

7.4.3 Viewing and filtering search results

Once search queries are generated using one or more query generation methods, fact-checkers can access and evaluate the corresponding search results on the *view-results* page. This page serves as a powerful tool, equipping fact-checkers to monitor, analyze, and track search results in a user-friendly and efficient manner. It provides a centralized hub where fact-checkers can conveniently view and manage multiple search queries

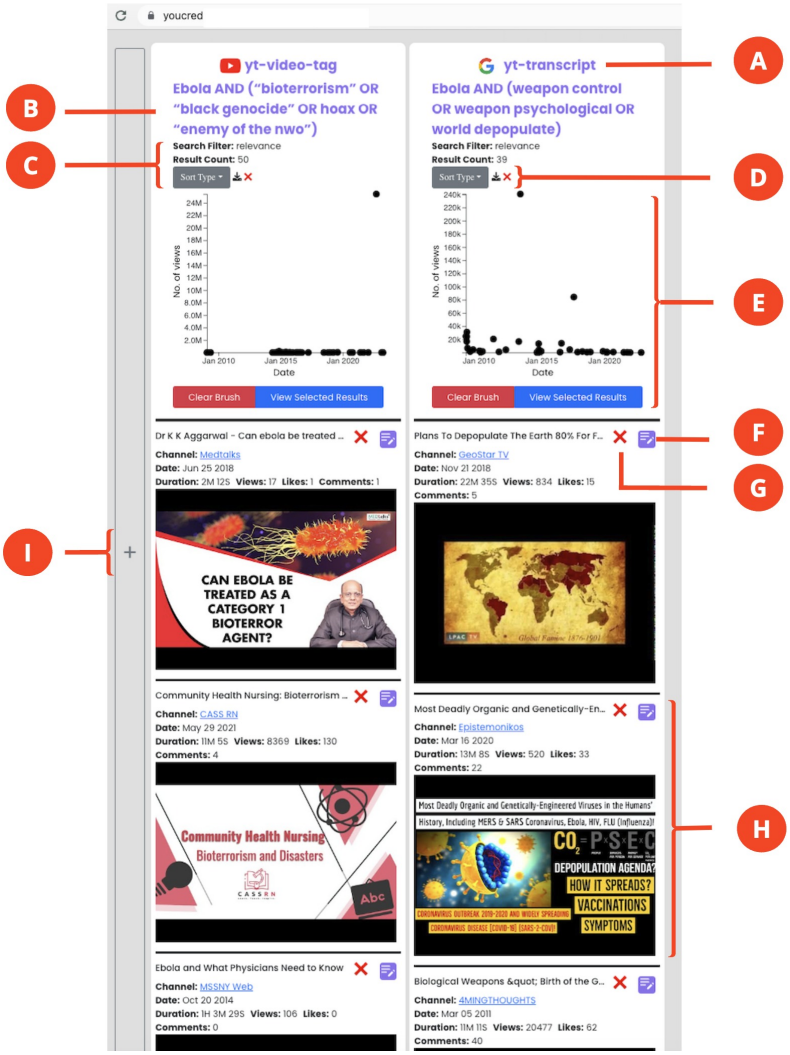


Figure 7.6: Snapshot of YouCred's *view-results* page consisting of multiple columns, with each column representing the search results of a specific query. The column header provides essential information such as the search query generation method (A), the search query itself (B), the applied sorting filter, and the count of search results (C). The page offers functionalities like downloading the search results as a CSV file and removing individual columns (D) as needed. Within each column, there is an interactive graph (E) that visualizes the engagement received by the search result videos and their publication dates. The page also includes sections dedicated to individual videos (H) representing each search result. These video sections provide important metadata such as the video title, channel name, upload date, views, likes, comments, and a thumbnail. If fact-checkers identify a potentially misinformative video, they can add it to the annotation database (F) for tracking and later fact-checking. Additionally, fact-checkers can utilize the block video functionality (G) to prevent a video from appearing in future search results.

simultaneously, enabling comprehensive evaluation and timely response to emerging trends and misinformation. The *view-results* page offers a range of capabilities and features designed to enhance fact-checkers' experience and effectiveness. These include the ability to customize search queries, an interactive graph for visualizing engagement metrics of the search results, access to search results metadata, video preview options, and other customization options to tailor the view and analysis of search results. Each of these features is described in more detail below, showcasing the rich functionality and flexibility provided by YouCred system.

7.4.3.1 Tracking Multiple Search Queries Simultaneously

The *view-results* page presents search results corresponding to the search queries as dedicated columns, ensuring a structured and intuitive layout. Fact-checkers can easily navigate between different search queries, enabling them to compare and evaluate various sets of search results efficiently. This organization enhances clarity and streamlines the fact-checking process. Figure 7.6 shows the snapshot of the *view-results* page with two columns, each corresponding to a different search query. The header of each column denotes the query generation method, followed by the search query, the search filter selected to sort the results, and the number of search results. The visually appealing and intuitive interface of YouCred ensures that fact-checkers can quickly scan and digest large volumes of videos without feeling overwhelmed. Fact-checkers can refresh and get the latest search results by simply getting double-clicking the query and pressing enter.

7.4.3.2 Customization through Editable Queries, sorting options, and addition of New Columns

Fact-checkers using YouCred have the flexibility to customize their search queries directly within the user interface. By double-clicking on a search query, they can easily edit it according to their specific needs. Once the editing is completed, simply pressing enter triggers YouCred to fetch and display search results for the modified query. Additionally, each column in the *view-results* page features a Sort Type drop-down button, allowing fact-checkers to further sort the fetched results based on criteria such as date, views, likes, or comments. This sorting feature empowers fact-checkers to focus on videos that are gaining engagement or target recently published content, enhancing their ability to prioritize fact-checking efforts.

YouCred also provides an option to remove a column by clicking on the ✕ button, enabling them to declutter the interface and focus on relevant search queries. Moreover, the search results within a column can be downloaded as a CSV file by clicking on the ⬇ button, facilitating easy data management and analysis. To expand their monitoring capabilities, YouCred enables fact-checkers to effortlessly add new columns dedicated to tracking additional search queries. By clicking on the plus symbol located on the left side of the page, fact-checkers can create new columns tailored to different search queries, allowing them to simultaneously track multiple topics or keywords in real time.

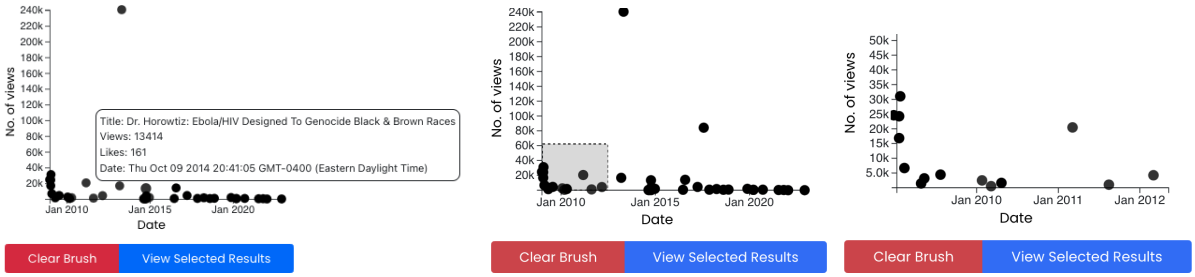


Figure 7.7: The *view-results* page in YouCred features an interactive, dynamic, and multifunctional scatter plot graph. This graph showcases the engagement received by the videos in the search results, represented on the y-axis, along with their respective dates of publication on the x-axis. (a) When hovering over a point on the graph, a text box displays detailed information about the video, including its title, engagement metrics such as likes and views, and the date of publication (Figure 7.7a). (b) Fact-checkers have the ability to select a specific cluster or area of interest within the graph (Figure 7.7b), (c) allowing them to zoom in and enabling a more focused analysis of selected videos (Figure 7.7c). The "View Selected Results" button filters the search results, displaying only the videos within the selected area, facilitating a more targeted evaluation. To revert back to the original graph view, fact-checkers can simply click the "Clear Brush" button, resetting the graph and allowing for further exploration and analysis.

7.4.3.3 Interactive Graphs for Engagement Metrics

Each column in the *view-results* page of YouCred features a dynamic, multifunctional, and interactive scatter plot graph. This graph showcases the engagement metrics on the y-axis and the date of publication of videos on the x-axis, providing valuable insights into the performance and timeline of the videos present in the search results. By default, the view count is displayed as the metric on the y-axis, providing an

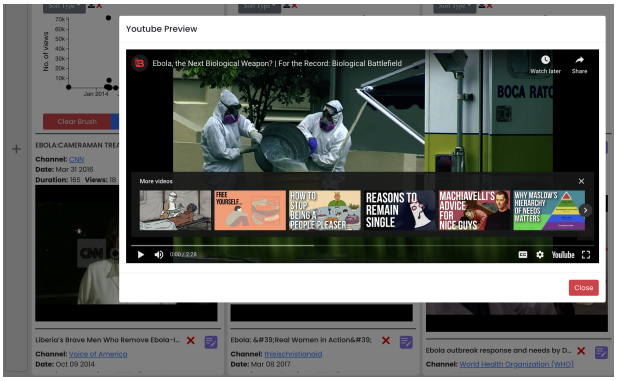


Figure 7.8: Snapshot of YouCred's preview mode. Fact-checkers can click on any video in the *view-results* page and can view the video in the system itself.

initial understanding of the popularity of the videos. However, fact-checkers have the flexibility to modify this metric based on their preferences. They can simply click on the sort-by drop-down button and select either the number of likes or comments as the desired metric.

Each point on the graph is a YouTube video present in the search result. Hovering on a point displays the video details including Title, like and view count, as well as date of publication of video. This information provides fact-checkers an overview of the video and its engagement metrics. The graph also aids in visualization and identification of any correlation or clustering of views based on the age of the videos or around a specific date range. To further explore specific clusters or areas of interest, fact-checkers can zoom in on the graph. By clicking and dragging the cursor around the desired region, they can filter out and focus solely on the zoomed-in version of the selected points, allowing for a more detailed analysis. In case any adjustments need to be made or additional points need to be included, users can easily reset the graph to its original state by clicking the "Clear Brush" button. Additionally, a convenient "View Selected Results" button allows users to exclusively view the videos within the selected area, aiding in focused analysis.

Overall, the dynamic and interactive scatter plot graph in YouCred's *view-results* page provides fact-checkers with a comprehensive and intuitive tool for visualizing engagement metrics and video publication dates. It enhances their ability to identify patterns, correlations, and clusters within the search results, facilitating efficient analysis and evaluation.

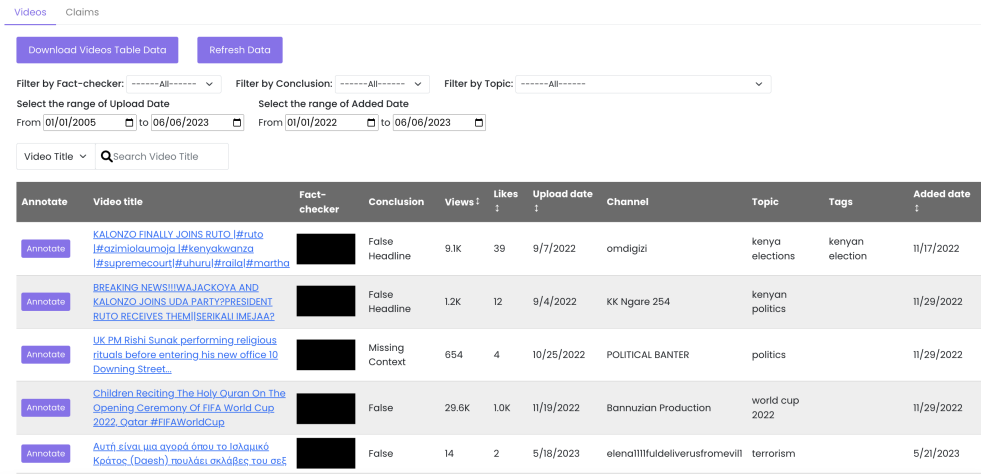


Figure 7.9: Figure shows the snapshot of YouCred's video annotation database. Fact-checkers add videos to this database while exploring the *view-results* page or directly from the browser extension. All videos have a corresponding annotate button which takes the fact-checkers to the video's annotation page. This database contains the video's title along with other metadata such as views, likes, upload date of video, channel, etc. Columns conclusion is populated once fact-checkers assign a veracity label to the video on the annotation page. Added date column denotes the date on which the video was added to the database. The page also provides a variety of search and filter options to find or view selected videos.

7.4.3.4 Video Metadata Sections

Below the interactive graph in each column, fact-checkers have access to detailed video sections. These sections present the search result videos along with their corresponding metadata, including the video title, channel name, upload date, views, likes, comments, and a thumbnail image. By clicking on the video thumbnail, fact-checkers can enter preview mode and view the video within the YouCred system.

Based on the video's metadata and preview, if fact-checkers identify the video as potentially misinformative, they have the option to add it to the annotation database (described in Section 7.5.1) for future fact-checking. To add the video to the annotation database, fact-checkers simply click on the button and can optionally add tags to aid in categorization. Additionally, fact-checkers have the ability to block a video by clicking on the ✕ button (refer to Section 7.4.2).

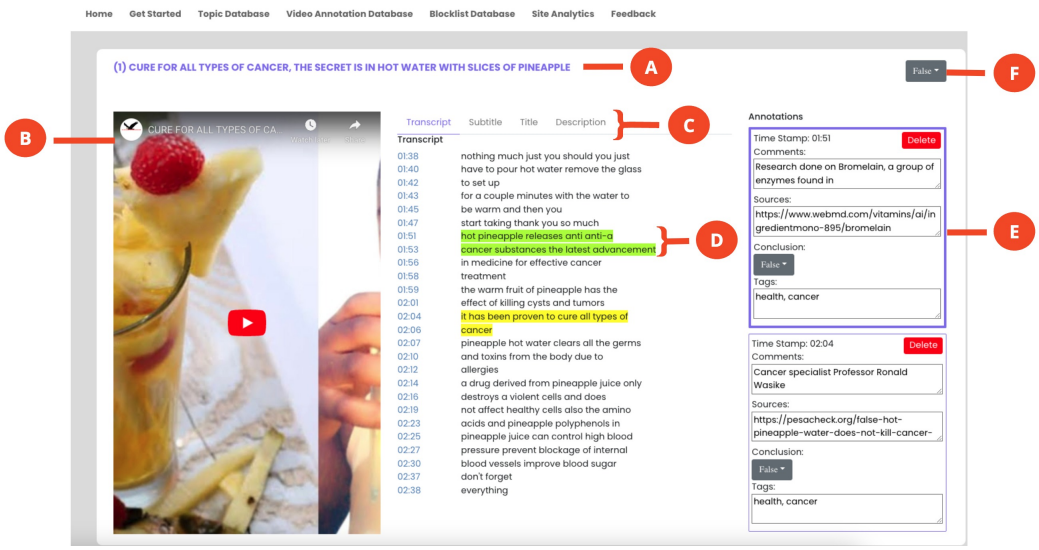


Figure 7.10: Figure showing YouCred's annotation page that streamlines and facilitates the credibility assessment process. The header corresponds to the video's title (A). The video is embedded towards the left side of the page (B) and the video's transcript, subtitles, title, and description are shown in the middle in separate tabs (C). Fact-checkers can highlight misinformative claims (D) in any tabs, add corresponding annotations (E) and also assign a veracity label to the video (F).

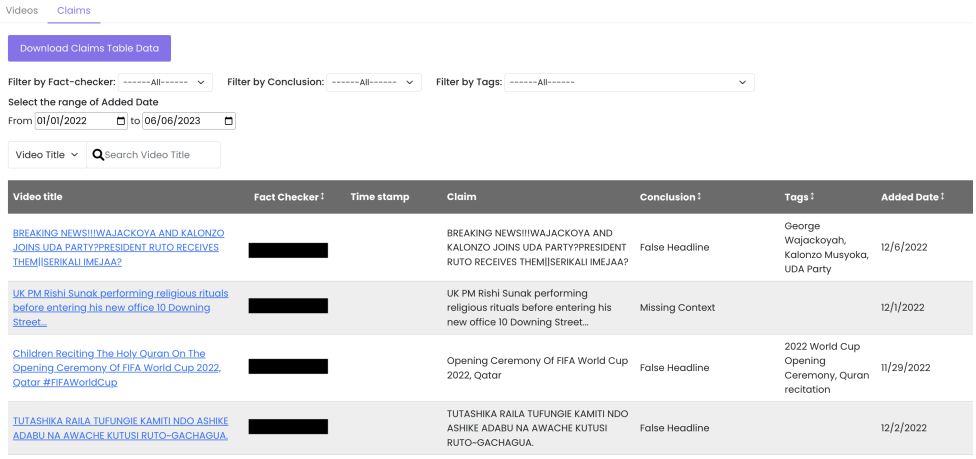


Figure 7.11: Snapshot of YouCred's claim database that stores entries for all the misinformative claims highlighted by fact-checkers in the videos that they annotated. The database shows the fact-checker name, the misinformative claim highlighted in the video, the veracity label of the claim, tags associated with the claim, and the date when the video was added to the annotation database.

7.5 Credibility Assessments

YouCred plays a crucial role in assisting fact-checkers with credibility assessments by providing a robust platform and valuable resources. First, YouCred offers an annotation database (Section 7.5.1) that allows fact-checkers to annotate videos, analyze transcripts, and assign veracity labels, enabling them to accurately assess the credibility of the information presented. Second, YouCred also curates a claims database (Section 7.5.3) that serves as a comprehensive repository, storing entries for misinformative claims highlighted by fact-checkers in annotated videos. It offers a snapshot of fact-checked claims, including details such as the fact-checkers name, the misinformative claim, the veracity label, associated tags, and the date of video addition. These centralized and structured databases enhance fact-checkers' ability to track, analyze, and combat misinformation effectively.

7.5.1 Video annotation database

Fact-checkers have multiple convenient methods to add videos to the annotation database: either while exploring the *view-results* page or directly from the browser extension. They have the flexibility to continuously add videos to the database and return to it later to annotate videos of their choice or prioritize important ones. Figure 7.9 shows a snapshot of the annotation database. Each video entry in the database includes an 'annotate' button, which, upon clicking, takes fact-checkers to the specific video's annotation page. This allows them to easily annotate the transcript in a new tab and provide accurate assessments. Similarly, clicking on the video title link opens the corresponding YouTube video page in a new tab, enabling fact-checkers to access additional context if needed.

The database ensures that essential metadata for each video is stored, including the video's title, number of views, likes, upload date, and channel information. The 'conclusion' column remains empty until fact-checkers assign a veracity label to the video on the annotation page, ensuring that conclusions are accurately reflected. The 'added date' column specifies the precise date when the video was added to the database, facilitating tracking and chronological organization of entries.

To facilitate efficient navigation and retrieval of specific videos, the page offers a range of search and filter options. Users can enter search keywords in the search box below, allowing for the filtering of videos based on desired criteria, such as keywords present in the video title. Furthermore, users can sort the table based on the selected

column by clicking on the headers of "Views," "Likes," and "Upload date." Clicking the headers toggles between not sorting, ascending sorting, descending sorting, and back to not sorting, enabling users to quickly arrange the table according to their preference.

7.5.2 Video annotation page

When fact-checkers want to annotate a specific video from the database, they can click on the 'annotate' button corresponding to that video, which directs them to the annotation page specifically designed for the selected video. Figure 7.10 provides a snapshot of the video annotation page. The title of the page corresponds to the title of the YouTube video. The page consists of four main components. The first component is an embedded YouTube video located on the left side of the page. The second component is the video profile area, which contains different tabs for the video's textual metadata, including the transcript, subtitle, title, and description. These components were selected by the fact-checkers as they play a crucial role in analyzing the veracity of the video. For example, Pesacheck assigns a veracity label of 'false headline' when the video title is misleading and does not accurately represent the actual content. These metadata components are computationally extracted using YouTube's API. If any of these components are missing, I provide an empty text box where fact-checkers can manually add and save the text. In the transcript tab, I display the text along with corresponding timestamps that are hyperlinked to specific time instances in the video.

The third component on the page is the annotations. Fact-checking is a multi-step process involving identifying misinformation claims, finding their sources, investigating their veracity, and writing fact-check reports. Throughout this process, fact-checkers curate information, consult experts, and make notes for themselves. Typically, fact-checkers use spreadsheets or text documents for this purpose, leaving comments containing information about sources, to-do tasks, and more. To streamline this process, I have incorporated annotations within the YouCred system. The annotation page offers a comprehensive solution for credibility assessments, eliminating the need for fact-checkers to switch between different documents and web pages. To begin the annotation process, a fact-checker can select the portion of text that contains a misinformation claim from any of the tabs. Once a text is selected, an annotation button appears, allowing fact-checkers to add annotations. The annotation appears as a text form, including a timestamp (in case the highlighted text is part of transcript) corresponding to the selected text. It also provides fields where fact-checkers can add

comments, sources, conclusions indicating the veracity label of the claim, and tags. Tags allow fact-checkers to assign specific labels or keywords to the claims which assists with categorization, organization and searchability. Once saved, the annotation appears on the right side of the page, and the corresponding text is highlighted in yellow. All fields in the text form are editable. Fact-checkers can easily locate their annotations by clicking on the highlighted text, which then becomes green. Similarly, clicking on the annotation block leads the fact-checker to the corresponding text. The fourth component of the page is the overall veracity label that fact-checkers can assign to the video. Note that each claim and the video as a whole can have different veracity labels.

7.5.3 Claims database

YouCred's claim database (Figure 7.11) serves as a comprehensive repository, capturing a snapshot of all the misinformative claims identified by fact-checkers within annotated videos. Each entry in the database contains essential information, including the name of the fact-checker who annotated the video, the specific misinformative claim identified, the veracity label assigned to the claim, relevant tags associated with the claim, and the date when the video was added to the annotation database. The database allows for the identification of recurring patterns or trends in misinformation. By analyzing the stored claims and associated tags, researchers can gain insights into common themes or topics prone to misinformation. This information can be used to develop targeted educational campaigns or policies to address specific areas of concern. Fact-checkers can refer to the database to determine if a similar claim has been fact-checked before. This feature helps to avoid duplicating efforts and allows fact-checkers to efficiently utilize their resources.

7.6 Evaluate stakeholders' acceptance

In early September 2022, I successfully completed all major features of YouCred and deployed the system, making it available to Pesacheck. To ensure effective adoption and utilization, I conducted two organization-wide training sessions for the fact-checking community at Pesacheck. The first training session took place on November 7, 2022, followed by the second session on March 13, 2023. Additionally, I created a demo video and provided comprehensive documentation of the system to support the organization

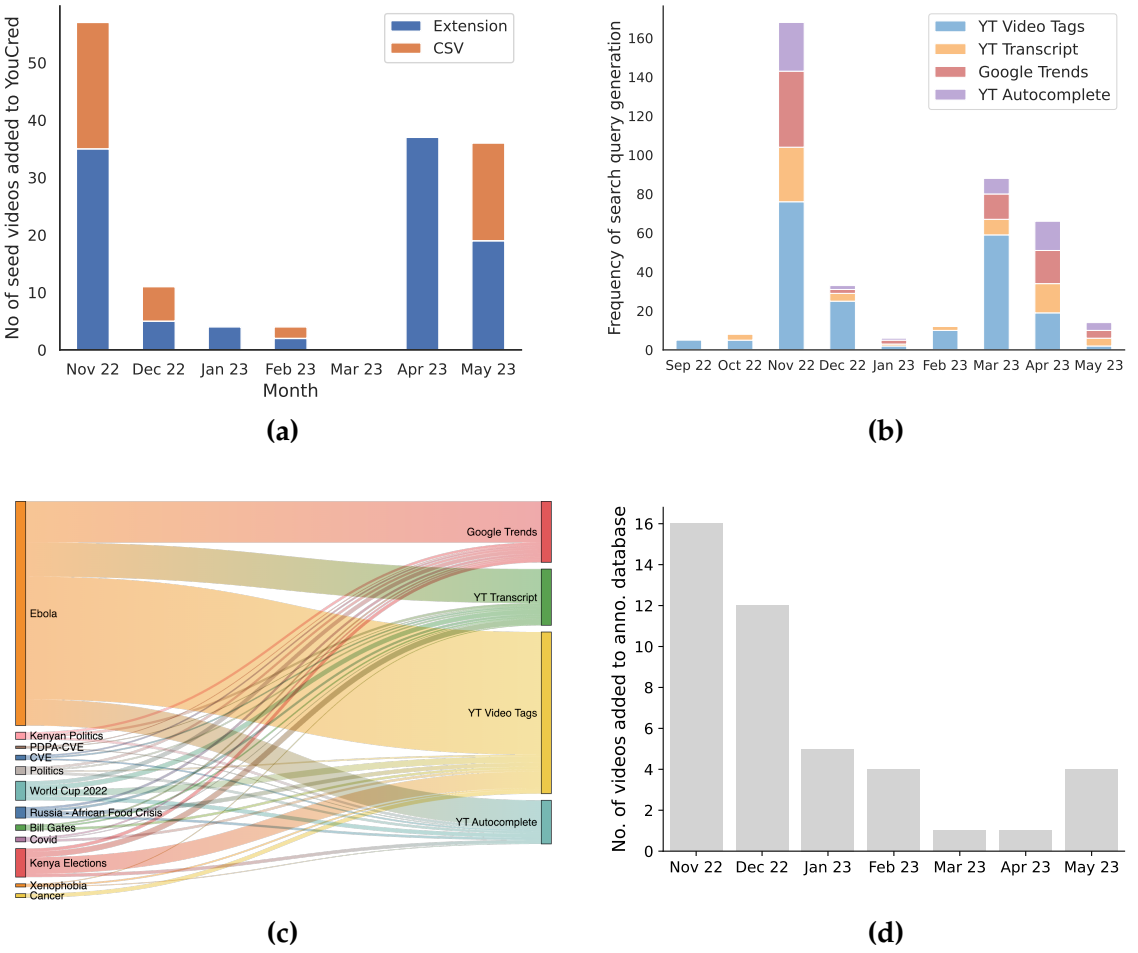


Figure 7.12: Figure (a) illustrates the number of seed videos added to YouCred through manual CSV uploads and the use of the 'YouTube-CSV-Helper' extension. Figure (b) presents the usage frequency of YouCred for generating search queries using the four proposed methods throughout the 9-month deployment period. Figure (c) provides an overview of the topics monitored using YouCred and the corresponding proportions of query generation methods utilized for each topic. Figure (d) illustrates the number of potentially misinformative videos added by fact-checkers to YouCred's annotation database.

in using YouCred effectively. Over a 9-month period, from September 2022 to May 2023, I closely monitored the usage of YouCred to assess its acceptance and impact within the organization. I employed two evaluation approaches for this assessment: tracking and analyzing the usage patterns of YouCred throughout its deployment and conducting semi-structured interviews with the fact-checking community to understand their overall perception and assessment of the tool's usefulness. In the following

sections, I delve into the details and findings of these two evaluation approaches.

7.6.1 Patterns of Usage over Time

YouCred's primary goals are to assist the organization in monitoring YouTube for misinformation discovery and credibility assessments. Therefore, I tracked the system's usage for search query generation and video annotation. It is important to note that following the deployment of YouCred, I observed a limited frequency of usage by the Pesacheck organization in the initial month. During a meeting with Pesacheck, I identified the manual creation and uploading of CSVs containing seed videos as a barrier to frequent usage. To address this issue, I developed and launched a helper extension in late October 2022. The usage of manual CSV uploads and the extension for adding new seed videos to the system is illustrated in Figure 7.12a[7]. As expected, the usage of the extension for providing seed videos was significantly higher compared to manual CSV uploads.

Throughout the 9-month deployment period, I tracked the frequency of search query generation using YouCred's four query generation methods, as depicted in Figure 7.12b. The usage frequency notably increased in November, aligning with the release of my extension. Among the query generation methods, video tags emerged as the most popular, followed by Google Trends and the transcript method. It is worth noting that Pesacheck's office was closed from December 15th to January 10th, and the system experienced downtime in January due to server issues. To address these challenges, I migrated YouCred from a local server to Microsoft Azure cloud services in mid-February, ensuring improved stability and accessibility. Both the office closure and server downtime impacted the usage of YouCred during that period.

Figure 7.12c showcases the various topics monitored on YouCred, along with the proportion of query generation methods used for each topic. I observed that the topic of Ebola received the highest level of monitoring, followed by the Kenyan elections in August 2022 and the FIFA World Cup held from July to August 2022. More recently, the system has been used to monitor discussions related to the Personal Data Protection Act (PDPA)[8,9] and Countering Violent Extremism (CVE) policies [10]. Figure 7.12d demonstrates the number of videos added to YouCred's annotation database

[7]It is important to note that I began recording the usage of the extension and CSV uploads in November, coinciding with the organization's adoption of the extension

[8]https://www.dataguidance.com/notes/south-africa-data-protection-overview

[9]https://www.dataguidance.com/jurisdiction/africa

[10]https://www.usaid.gov/policy/countering-violent-extremism

for credibility assessments. Fact-checkers have added 45 videos about various topics mentioned in Figure 7.12c for credibility assessments. Overall, the quantitative analysis indicates consistent usage of YouCred during the 9-month deployment, despite the temporary downtime and the few bugs that I addressed after the initial deployment.

7.6.2 Semi-structured interviews

To gain a comprehensive understanding of YouCred's adaptation and usage, I complemented the quantitative evaluation with semi-structured interviews with the fact-checking team. In this section, I provide a brief overview of the interview protocol and summarize the key findings derived from these interviews.

7.6.2.1 Participants and Interview Procedure

I utilized a purposive sampling technique [255] to select participants for the semi-structured interviews. I reached out to both regular users of YouCred, identified from the video annotation database and extension log, as well as individuals who actively participated in the design process. The sample consisted of six stakeholders from Pesacheck (referred to as I1-I6), including fact-checkers and advocates, all of whom had utilized YouCred during its deployment phase. Notably, four participants were regular attendees in the design process meetings.

The semi-structured interviews were conducted remotely over Zoom by the first author, with the consent of the participants. All six interviews lasted approximately 60 minutes and took place between June and July 2023. Throughout the interviews, the first author made detailed notes based on observations.

7.6.2.2 Interview Protocol and Data Analysis

I began the interviews by asking fact-checkers to provide a brief explanation of their previous methods for monitoring the YouTube platform before the deployment of YouCred. This allowed the fact-checkers to elaborate on the limitations they faced with their existing methods and set the stage for discussing how YouCred addressed those limitations. I then focused on how the participants integrated YouCred into their fact-checking workflow and explored specific scenarios and topics where they had utilized the tool. To gain deeper insights into their usage, I selected one of the topics discussed and requested the participants to share their screen and guide me through their process of searching for that topic on YouTube. Subsequently, I asked

the participants to demonstrate how they monitored the same topic using YouCred. I encouraged them to walk me through their step-by-step process and highlight the differences between directly searching on YouTube and utilizing YouCred. Throughout this walk-through, I inquired about the benefits and usefulness of each functionality provided by YouCred. I also prompted the participants to evaluate and discuss the advantages and usefulness of each query generation method available in YouCred.

In addition, I conducted an exercise where I displayed the YouTube search results interface and the YouCred *view-results* page side by side on the screen, enabling direct comparisons. I specifically asked the participants to share their insights on how the two interfaces differed and how these differences influenced their fact-checking process. Throughout the interview, I delved into YouCred's effectiveness and discussed how the system aligns with their specific needs. I also actively encouraged the interviewees to provide suggestions for improving YouCred and discuss any limitations they may have encountered during their usage.

The first author transcribed all the interviews and notes and analyzed them using an iterative qualitative thematic approach. This method allowed me to uncover significant insights and patterns emerging from the participants' responses.

7.6.2.3 Findings

By analyzing interview transcripts and notes, I gained valuable insights into the utility of YouCred. In this section, I delve into the key findings from the interviews and also discuss potential areas for improvement as suggested by the participants.

Fact-checkers have integrated YouCred into their day-to-day workflow. Fact-checkers revealed that YouCred "*comes in handy because of the lack of available free tools available to look for misinformation online, especially on Youtube*" (I3). They revealed various ways in which they have integrated YouCred into their workflow. For example, I3 revealed that most of the time "*viral misinformation will go viral on all social media sites*", so whenever they find a misinformative video, "*even if it's on Facebook*" they "*try to feed it into the YouCred system*". I5 revealed that they have also used YouCred to monitor some persistently circulating misinformation around health-related topics like COVID and Ebola.

YouCred provides rich and meaningful query suggestions. YouCred's query generation methods proved to be a valuable resource for fact-checkers providing "*a pool*

of information that would otherwise not be available if [they] were generating queries manually" (I4). Fact-checkers also appreciated the flexibility to customize their searches on YouCred. As one participant expressed, "*I like YouCred suggesting keywords. So yeah, and also being able to edit those keywords and also adding your own, because that refines your search a lot and leads to a higher possibility of getting what you want with the YouCred tool but with the YouTube keyword search, it's a hit or miss*" (I3).

Fact-checkers also highlighted the effectiveness of "*YouCred's integration with Google Trends*" which consistently delivered "*high-quality search results for them*" (I3). In addition, the use of video tags within YouCred unveiled additional insights into the strategies employed by content creators to boost the popularity of misinformative videos. For instance, I4 observed that "*content creators use the names of celebrities as tags on their YouTube video so that they can trend*".

YouCred facilitates modular assessment of videos. Participants revealed that they mostly write short-form fact-check reports which debunk an individual claim. However, if the video contains multiple false claims, they'll either write multiple short-form reports for each misleading claim or one long-form article debunking all claims. In both scenarios, fact-checkers have found YouCred's annotation interface to be incredibly valuable, enabling them to break down the video into distinct claims and conduct focused investigations.

> "When we look at multiple items within the video that might be false, YouCred allows us to break down the video into multiple claims and investigate each section" - I1

Fact-checkers use YouCred's claim database as an informative resource. Fact checkers' consider YouCred's claim database as a powerful resource that goes beyond simply cataloging misinformative claims. For example, P4 suggested that the tool helps them avoid duplication of effort since "*you can realize that you know, someone else has had really added the video [in the claim's database]..and is working on it*". In addition, they find the addition of channel details on the page useful as "*it will let you know the frequent spreaders of misinformation...So you can actually go to their channel and see, if have they shared any more misinformation*". Additionally, another participant expressed that they are now able to explore a collection of videos related to a specific topic and claim, enhancing their understanding of the claim's prevalence and context.

> "[Tags] on claim database provides an option to see a cluster of videos within one topic, so I believe, probably you will see this video contains this claim And

> you'd probably find another video that contains the same claim so on the claims database you're able to quickly see which the number of videos that are actually speaking about one specific claim, so I think it's a fantastic feature" - I2

Youcred enhances the efficiency of fact-checking workflow. Participants revealed that using "*YouCred tool, [they] could spend less time getting misinformation as opposed to when they were searching on YouTube directly*" (I5). They also talked about several features that have played a vital role in improving the efficiency of their fact-checking workflow. According to I2, the preview mode on the *view-results* page is a valuable tool that allows fact-checkers to do "*quickly preview [the video] on the same page and decide if they even want to annotate a video in the first place without going to Youtube.*". I3 highlights the usefulness of the ability to track multiple search queries on the *view-results* page by stating, "*they now have everything on one page, and we're able to monitor everything at the same time...instead of doing multiple searches in several tabs..it saves you more time*". Additionally, the interactive graphs in Youcred have also garnered positive feedback from participants. As I3 expressed, "*The interactive graphs are very helpful. You're able to only focus on what you need. You can just select a particular period, and also even the number of views*".

Possible future improvements and enhancements. During the discussions, participants highlighted the need for continuous improvement in systems like YouCred. They made valuable suggestions to enhance the system and meet the evolving needs of fact-checkers. One common complaint was the limited YouTube API quota, which at times hindered the seamless use of the system. Participants mentioned that they desire the ability to track multiple topics simultaneously, as well as the convenience of viewing all suggested queries directly in the drop-down menu on the *view-results* page, eliminating the need to navigate back to the methods page to select new keywords. Some participants requested us to further refine search results using AI to only show videos that could contain problematic content. Looking towards the future, one fact-checker envisioned the claims database to be integrated with the YouTube platform itself. I4 proposed that YouTube could extract videos and their associated misleading claims, along with timestamps, and display a disclaimer indicating the presence of misinformation. The aim would be to instill hesitation in viewers when considering sharing such videos. I4 considers this intervention as a potential strategy to promote responsible information sharing and combat the spread of misinformation.

7.7 Discussion

In this study, I present YouCred, a fact-checking system designed and built to assist fact-checkers in monitoring the YouTube platform. The formative study highlighted the manual nature of the platform monitoring process and the lack of tools to aid credibility assessments of YouTube videos, leading me to develop YouCred as a solution. YouCred serves as a successful case study, demonstrating the importance of continued dialogue and collaboration with fact-checking organizations in designing systems that have a tangible impact on real-world fact-checking practices. In this section, I reflect on the design implications of my work and discuss, how we need to think about the maintenance of socio-technical systems beyond the initial deployment period.

7.7.1 Design Implications

7.7.1.1 Bridging design-reality gap in design of fact-checking systems

The design of fact-checking systems often falls short of making a significant impact on real-world fact-checking practices due to a lack of incorporation of insights and needs from the various stakeholder groups involved in the process [317]. Scholars have highlighted this gap, emphasizing the importance of involving key stakeholders to ensure the relevance and effectiveness of such systems [317]. In my work, I aimed to address this issue by adopting participatory design methods that actively engaged key stakeholders from fact-checking organizations throughout the design process. This approach allowed me to foster meaningful dialogue, gain valuable insights, and better understand the perspectives and concerns of fact-checkers.

By involving fact-checkers in the design of YouCred, I was able to create a system that truly catered to their needs. The participatory design process facilitated collaborative decision-making, where fact-checkers' expertise and experiences directly influenced the system's functionalities and features. Through continual engagement, I built a deep understanding of the challenges they faced, such as the manual nature of monitoring the YouTube platform and the lack of tools for credibility assessments of videos. As a result, YouCred was purposefully designed to address these specific challenges and provide fact-checkers with effective tools for their daily work. This approach emphasizes the importance of collaboration and knowledge exchange between researchers and practitioners, ensuring that the resulting system is not only technically robust but also practical and impactful in supporting fact-checking efforts. Moving forward, such participatory approaches can serve as a valuable framework

for designing socio-technical systems that address the needs of various stakeholder groups involved in complex real-world domains.

7.7.1.2 Imbibing values desired by fact-checkers in design of fact-checking systems

Fact-checking is a socio-technical process that goes beyond the mere application of technology [317]. To ensure the successful adoption and effectiveness of fact-checking systems, it is crucial to incorporate the values and preferences of fact-checkers into the design [317]. The formative interviews with fact-checkers revealed their desire for agency and human involvement in the design of these systems. This value signifies the importance of integrating human expertise and judgment into the decision-making processes of fact-checking systems. Building upon these insights, I developed YouCred to provide fact-checkers with a sense of control and agency over their work. One way this was achieved was by empowering fact-checkers with the ability to customize their search queries on YouTube. The system allows fact-checkers to add, modify, or remove terms from the generated search queries, or even create their own queries to monitor relevant content. By granting this level of flexibility, fact-checkers can tailor their searches to align with their specific needs and interests, enhancing the effectiveness and efficiency of their fact-checking efforts. Imbibing the value ensures that fact-checkers are not merely passive users of the system but active participants in shaping its functionality and outcomes.

7.7.2 Maintainability of socio-technical systems

Deploying a fact-checking system in the real world is a complex undertaking that goes beyond the initial development and deployment phase. It requires ongoing maintenance and continuous improvement to ensure its effectiveness and relevance in the ever-evolving technological and information landscape. As I deployed YouCred, I encountered various challenges, including scalability, concurrency, and the need for real-time bug handling to ensure uninterrupted system usage. I also realized that the needs of fact-checkers are not static but evolve over time, emphasizing the need for constant adaptation and evolution of the fact-checking system. Simply building and deploying a system is insufficient; active maintenance and enhancement are essential to meet the evolving requirements of fact-checkers. This underscores the significance of investing resources and efforts into the ongoing maintenance of deployed systems.

In recent literature, scholars have raised concerns about sustaining maintainability efforts after the initial project funding ends [240]. Maintenance and repair activities have been recognized as overlooked yet essential aspects of socio-technical initiatives, encompassing creativity, innovation, knowledge, power dynamics, and ethics of care [211]. Efforts have been made to study the challenges of sustaining and maintaining the outcomes of Human-Computer Interaction (HCI) design and systems projects beyond their runtime and beyond the researchers' role within the project context [130, 211, 212, 240]. For instance, Meurer et al. suggest the importance of nurturing a sense of ownership among research participants from the project's outset and developing their technological capabilities [377]. However, the feasibility of such approaches varies greatly depending on contextual factors, including the technical and human resources available to different researchers and stakeholders for maintaining the designed technological artifacts. Hence, scholars argue that active support from funding bodies is crucial to encourage and enable researchers by allocating time and resources specifically for maintenance activities, ensuring the continuity of these efforts beyond the project's duration [240].

Maintaining a real-world deployed system also requires a collaborative effort between researchers, developers, and other stakeholder groups. While, there is an active research area in software engineering focusing on developing strategies, frameworks, and tools that facilitate the maintainability of software systems [253, 349, 403], we also need to explore the collaborative efforts in the maintenance and evolution of socio-technical systems. Additionally, a significant investment of time and effort is dedicated to cultivating and sustaining social relationships with stakeholder groups throughout the development and maintenance of such collaborative socio-technical systems projects. Recognizing the importance of these relationships, even beyond the initial stages, is crucial. However, these efforts often fall outside the scope of traditional design or research activities. [240]. Encouraging and supporting such endeavors would not only ensures the long-term sustainability of collaborative initiatives but also enhance their overall impact.

7.8 Limitations and Opportunities

While YouCred has demonstrated significant potential in assisting fact-checkers in combating online misinformation on YouTube, there are limitations to address and numerous opportunities for further development and expansion. I discuss a few below.

- Limited YouTube API Quota: The current version of YouCred relies on the YouTube API to extract search results and video metadata. However, the API has a restricted quota, limited to 10,000 requests per day. Fact-checkers often quickly max out their individual API quota. To address this problem, I filled out an official form requesting increased quota limits but did not receive any response. However, I am hopeful about the opportunities provided by the YouTube Researcher Program[11], which offers expanded access to the YouTube API. I am actively applying to join this program to address the quota limitations.
- Dependency on YouTube API for Transcripts: YouCred currently depends on the YouTube API to extract video transcripts, which are displayed on the video annotation page. However, this approach limits the availability of transcripts, as not all videos have them. To overcome this limitation, I plan to explore video-to-text conversion tools in the future. By utilizing these tools, I can obtain transcripts for videos that do not have them available, further enhancing the annotation capabilities of YouCred.
- Single-Topic Monitoring: The current version of YouCred allows fact-checkers to monitor YouTube for one topic at a time. This limitation hinders their ability to track multiple topics simultaneously. To address this limitation and provide a more comprehensive monitoring solution, I plan to expand and improve YouCred. My goal is to develop functionality that enables fact-checkers to easily track and monitor multiple topics concurrently, enhancing their efficiency and effectiveness in combating misinformation.
- Deployment to Other Fact-Checking Organizations: As part of my future plans, I aim to deploy YouCred to other fact-checking organizations. By sharing this tool with different organizations, we can create a network where resources, such as the claims database, can be shared among them. This collaborative approach not only enhances fact-checking efforts but also provides an opportunity for researchers to study the misinformation landscape across different countries, furthering our understanding of this global challenge.
- Integration with Tiplines and User Reporting: Many fact-checking organizations have established tiplines where users can report potentially misinformative content. To leverage user contributions and expand the reach of YouCred, I plan to integrate the system with these tiplines. By encouraging users to report search queries that lead them to misinformation online, I can enhance the effectiveness

[11] https://research.youtube/

and scope of YouCred in identifying and addressing misinformation on YouTube.

- Dependency on seed videos: Youcred requires fact-checkers to curate a few seed videos in order to further monitor that topic. This curation process can be time-consuming and labor-intensive. Integrating the system with user tiplines is one way to address this issue. User-reported videos can not only provide valuable insights into the problematic topics that users are exposed to, but they can serve as alternative seed videos.
- Refining search results: As part of my future plans, I aim to enhance the display of search results in YouCred. As fact-checkers fact-check more videos about a topic, the system can utilize machine learning algorithms to identify and prioritize search results with a higher probability of containing misinformation. One potential way is to look for linguistic signals in the videos' comments to determine if the video could potentially contain misinformation [217].

7.9 Conclusion

In this study, I build a fact-checking system that helps fact-checkers with misinformation discovery and credibility assessments on YouTube. The system is a result of a 2-year collaboration with Pesacheck—Africa's largest indigenous fact-checking organization. Throughout the entire development process, I actively engaged the fact-checking team, involving them in requirement elicitation, design iterations, and evaluation phases. To evaluate the effectiveness and user acceptance of YouCred, I deployed the system at Pesacheck and monitored its usage for a duration of nine months. In addition, I conducted follow-up semi-structured interviews with the fact-checkers to gather their insights and feedback. The results of this comprehensive evaluation revealed a positive reception of YouCred within the fact-checking community. The fact-checkers acknowledged the system's utility and found it valuable in their daily fact-checking endeavors. Overall, this work validates the effectiveness of participatory approaches in designing fact-checking systems that effectively meet the needs and expectations of fact-checkers.

CHAPTER 8

FUTURE WORK AND CONCLUSION

In the current digital landscape, people increasingly turn to search engines and social media platforms for news and information. However, the content presented by online platforms is not always reliable. The information can be biased, inaccurate, or even misleading. This situation is further compounded by the risk of users getting trapped in filter bubbles that expose them to even more problematic content. In my thesis work, I delve into this issue, which I term "algorithmically curated problematic content," with a specific focus on misinformation. I examine how algorithms contribute to presenting and amplifying misleading content and design defenses against such content presented by platforms through three distinct research threads.

In the first research thread, I developed audit methodologies to assess how user attributes and activities influence the extent of misinformation that surfaces in search results and recommendations. Applying these methodologies, I conducted a comprehensive series of meticulously controlled experiments on various social media search interfaces, including YouTube (Chapter 3 and 4) and Amazon (Chapter 5). I found that these platforms amplify the misinformative content to users under certain conditions. I also identified vulnerable user populations who could be targets for certain misinformative topics on online platforms. Given the significant influence of online platforms and the lack of universal policies against harmful online content, I advocate for a shift in responsibility from users to platforms. I believe platforms must take a more proactive role in ensuring the accuracy and reliability of the information presented to their users. This may involve rethinking traditional recommendation algorithms and

treating topics that directly impact users' well-being, health, and happiness with extra scrutiny and ensuring high-quality searches and recommendations for them.

The second thread of my research deep dives into the ways we can address the problem of online misinformation that gets surfaced by online platforms. For this work, I turned to fact-checking organizations that are constantly monitoring online platforms to determine the veracity of potentially dubious claims. Including the voices of those battling misinformation is essential, as they offer insights into real-world challenges and needs. Equally important is identifying the human stakeholders within these organizations and their roles. This knowledge empowers us to support all facets of the fact-checking process, both visible and invisible. Consequently, I interviewed individuals from fact-checking organizations across four continents. Through this study, I deep-dived into the process of online fact-checking by foregrounding the human and technological infrastructures of the fact-checking process (Chapter 6). I also unraveled the barriers to fact-checking online misinformation. Based on my findings, I propose that improving the quality of fact-checking necessitates systematic changes within the civic, informational, and technological contexts. Such changes are essential for a comprehensive approach to mitigating misinformation effectively. This research provides a road map for future investigations in fact-checking space, offering valuable insights that can aid and enhance the endeavors of fact-checking organizations.

In the final thread of my dissertation, I leverage the insights gained from my previous research to design and build a fact-checking system (Chapter 7). The need for the system emerged from my interview study with fact-checking organizations, which revealed the challenges in monitoring video search engines like YouTube. Unlike platforms such as Twitter and Facebook, YouTube lacks trending topics or public groups for information sharing. Therefore, fact-checkers end up manually searching the platform by crafting queries based on guesswork. I solve this problem by building the YouCred fact-checking system. The major contribution of this project is the use of participatory methods where fact-checking stakeholders were included in all stages of system development, from requirement elicitation to design, and evaluation. I build the solution based on fact-checkers existing knowledge and processes to ensure the system's long-term sustainability. This work demonstrates that for effective solutions that can impact fact-checking processes in the real world, collaboration with fact-checkers and their inclusion in the design process is imperative.

In conclusion, my thesis documents an extensive investigation into identifying,

measuring, and defending against algorithmically curated online misinformation. My approach adopts a multifaceted perspective, scrutinizing online misinformation through three distinct lenses: algorithms, fact-checking infrastructure, and design. Through this multifaceted approach and close collaboration with diverse stakeholders, I have been successful in developing effective solutions that can make a real difference in the fight against online misinformation. My research contributes not only to academic discourse but also offers practical insights into designing better technology and policies to mitigate the harmful effects of online misinformation.

8.1 Future work

Each research thread of my dissertation paves the way for new avenues of exploration. In this section, I delve into the potential future directions of my work. First, I discuss promising avenues that could be pursued within the realm of algorithmic audit research (Section 8.1.1). Second, I propose strategies to counter algorithmic harm by designing for algorithmic awareness (Section 8.1.2). I propose to seek ways to incorporate algorithmic literacy early in the education system by piloting interventions in university courses. I also propose designing frameworks to generate explanations of how algorithms work and impact users. While awareness about how the algorithms function and how they can cause harm is important, we also need pathways to revert undesirable algorithmic behaviors. Thus, my third proposed work focuses on designing for algorithmic recourse that allows users to change undesirable algorithmic decisions (Section 8.1.3). Finally, my dissertation research mostly focused on US-centric misinformation in the English language. I am eager to extend my investigations to the Global South region (Section 8.1.4). This future research trajectory would enable me to better understand the nuances of the misinformation phenomenon within diverse cultural contexts and languages.

8.1.1 Exploring New Horizons in Algorithmic Audit research

I think there is a lot to be done in the algorithmic audits space. A recurring pattern in most audit studies (including my own) is the limited timeframe—they tend to assess a platform over a short span. In my view, there's immense value in embracing continuous audits that unfold over months or even years. This prolonged perspective could unveil how platforms' policies and algorithmic behavior evolve over time. Fur-

thermore, certain platforms like Spotify, ephemeral content generators like Snapchat, and non-Western search engines like Yandex have received limited attention in audit research. Additionally, most audits are focused on the Global North regions. Auditing algorithms in the Global South and even conducting cross-country or cross-continent audits could provide invaluable insights into regional variations in algorithmic behavior in two different parts of the world. Multi-platform audits also present an unexplored avenue. Such audits would scrutinize and evaluate the impact, fairness, and overall behavior of algorithms across various digital spaces. The goal would be to gain insights into how these algorithms curate content, make recommendations, and shape user experiences on different platforms. Consider an example, few extremist groups have a presence on both Twitter and Facebook [313]. But there has been no study to test and compare the effects of following these accounts on both platforms. Multi platforms audits can help understand which platform exacerbates the effect of a problematic user action and to what extent.

The last year of my Ph.D. witnessed an explosive rise and interest in generative AI powered by large language models. Generative AI is now being integrated with search engines like Google and Bing and has brought forth concerns about its potential to propagate harm and amplify biases. I believe there is a huge potential in auditing research space to investigate these platforms for bias. This includes evaluating the factual accuracy of their responses to queries about important topics, as well as understanding their behavior when presented with prompts that are either multilingual or in non-western languages.

8.1.2 Designing for algorithmic literacy and awareness.

One of the first steps towards redressing algorithmic harm is to make people aware of the presence of algorithmic systems which would then allow users to question the outcomes of the algorithmic system. I want to understand how can we incorporate algorithmic literacy early in the education system by piloting interventions in high school courses as well as university courses in non-IT degrees. I want the interventions to educate people not only about the presence of algorithms but also about how their actions can impact algorithmic behavior. Next, I want to create frameworks to generate clear, meaningful, and useful explanations of how algorithms work and impact users. The vision of this work closely aligns with the White House's AI Bill of Rights and I'll be applying for federal funding for the same. I've already made headway in this line of inquiry [316]. I have proposed design interventions that act as 'decision aids' to

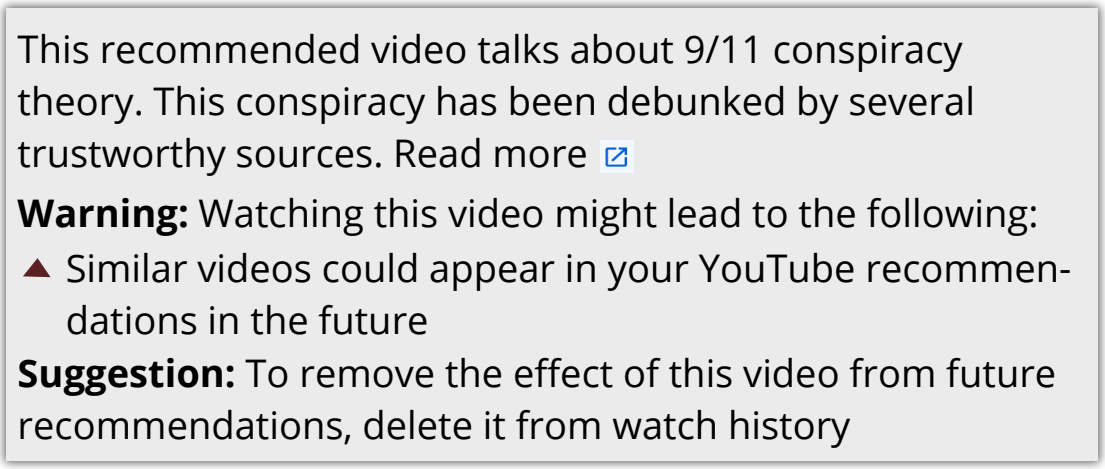


Figure 8.1: When a conspiratorial video gets recommended on a user's YouTube homepage, the user is warned about the consequences of watching the video on future video recommendations.

users when algorithms expose them to problematic content [316]. My design presents users with facts (what happened) accompanied by forewarnings (what could happen) to convey the potential risks of action in a comprehensible manner. Figure 8.1 shows a mock-up of my design intervention.

8.1.3 Designing for algorithmic recourse.

Awareness of the presence of an algorithm empowers users to question or contest the algorithmic decision. What does a user do when they encounter an undesirable algorithmic outcome? The field of machine learning (ML) has introduced the concept of recourse which is defined as the ability to change decisions of an ML model by changing input variables. I want to take a human-centered approach to recourse. I plan to understand users' needs for recourse from the online platforms and investigate the current systems for recourse settings (e.g. setting to indicate that a user is not interested in content from a particular source). I plan to redesign existing online platforms for algorithmic recourse so that users have pathways to change the algorithmic decisions by performing particular actions on the system.

8.1.4 Studying misinformation, fact-checking, and algorithmic impact beyond the US.

While misinformation is a global crisis, measures to combat it vary with respect to culture, geographic location, language, etc. Most of the academic research, to date, has primarily focused on combating misinformation in Western countries, while not addressing the phenomenon in the Global South. How is fact-checking practiced in the Global South? What are the various barriers to the fact-checking process in the Global South? How can we design tools and technology to assist the stakeholders in the fact-checking process in a resource-constrained context of the Global South region? I want to answer these questions by first interviewing various stakeholder groups involved in the fact-checking process in the Global South and then using the insights to build tools for them. I strongly believe that this research would help provide a global perspective on misinformation and fact-checking. I am already collaborating with several fact-checking organizations in the Global South to make headway in this research direction.

In a separate line of inquiry, I want to conduct systematic, effective, and ethical audits on various online platforms for problematic content such as hate speech, extremism, and conspiracy theories specific to the Global South countries. Many online platforms have content moderation policies for the aforementioned problematic content. However, it's not known whether these policies are being applied uniformly in the Global North and Global South region. There also have been reports of inconsistent support for misinformation and hate speech in non-English languages on social media platforms, which disproportionately affects those in the Global South where such content spreads through local regional languages. I want to investigate how problematic content in non-English languages gets surfaced online and in turn understand, how effectively online platforms have enacted their policies in the Global South.

BIBLIOGRAPHY

[1] *Amazon slammed for promoting false covid cures and anti-vaccine claims : Npr*.
https://www.npr.org/2021/09/09/1035559330/democrats-slam-amazon-for-promoting-false-covid-cures-and-anti-vaccine-claims.
(Accessed on 12/28/2022).

[2] *Fighting coronavirus misinformation and disinformation - center for american progress*.
https://www.americanprogress.org/article/fighting-coronavirus-misinformation-disinformation/.
(Accessed on 01/10/2023).

[3] *How google's search algorithm spreads false information with a rightwing bias | google | the guardian*.
https://www.theguardian.com/technology/2016/dec/16/google-autocomplete-rightwing-bias-algorithm-political-propaganda.
(Accessed on 12/28/2022).

[4] *Youtube is still struggling to rein in its recommendation algorithm*.
https://www.buzzfeednews.com/article/carolineodonovan/down-youtubes-recommendation-rabbithole.
(Accessed on 12/28/2022).

[5] *Youtube more likely to recommend election-fraud content to those skeptical of the 2020 election: study – the hill*.
https://thehill.com/changing-america/enrichment/arts-culture/3625989-youtube-more-likely-to-recommend-election-fraud-content-to-those-skeptical-of-the-2020-election-study/.
(Accessed on 12/28/2022).

[6] *List of conspiracy theories*, (2019).

[7] *Google search help*, (2020).
https://support.google.com/websearch/answer/9281931?hl=en.

[8] *3 challenges of integrating heterogeneous data sources - dzone integration*.
https://dzone.com/articles/3-challenges-of-integrating-heterogeneous-data-sou, August 2021.
(Accessed on 08/03/2021).

[9] *Challenges of integrating heterogeneous data sources - dataversity*.
https://www.dataversity.net/challenges-of-integrating-heterogeneous-data-sources/, August 2021.
(Accessed on 08/03/2021).

[10] *A common data model for europe? - why? which? how? - workshop report*.
https://www.ema.europa.eu/en/documents/report/common-data-model-europe-why-which-how-workshop-report_en.pdf, Aug 2021.
(Accessed on 08/03/2021).

[11] *A common data model in europe? – why? which? how? | european medicines agency*.
https://www.ema.europa.eu/en/events/common-data-model-europe-why-which-how, Aug 2021.
(Accessed on 08/03/2021).

[12] *Der spiegel | online-nachrichten*.
https://www.spiegel.de/consent-a-?targetUrl=https%3A%2F%2Fwww.spiegel.de%2Finternational%2F&ref=https%3A%2F%2Fwww.google.com%2F, September 2021.
(Accessed on 09/14/2021).

[13] *dpa: en*.
https://www.dpa.com/en/, August 2021.
(Accessed on 08/17/2021).

[14] *Fact check*.
https://www.indiatoday.in/fact-check, August 2021.
(Accessed on 08/17/2021).

[15] *Portada · maldita.es - periodismo para que no te la cuelen*.
https://maldita.es/, August 2021.
(Accessed on 08/17/2021).

[16] *Presentación de powerpoint*.
https://maldita.es/uploads/public/docs/barometro_desinformacion_parte_1.pdf, Aug 2021.
(Accessed on 08/25/2021).

[17] *Search results - lexisnexis*.
https://www.lexisnexis.com/en-us/search.page, April 2021.
(Accessed on 04/15/2021).

[18] *Uganda in crisis – ancir's ilab*.
https://investigate.africa/reports/uganda-in-crisis/, Aug 2021.
(Accessed on 08/25/2021).

[19] *Bringing fact check information to google images*.
https://blog.google/products/search/bringing-fact-check-information-google-images/, January 2022.
(Accessed on 01/13/2022).

[20] *Debunk bot (@debunkbotafrica) / twitter*.
https://twitter.com/debunkbotafrica, Jan 2022.
(Accessed on 01/12/2022).

[21] *Facebook takes down fact-check of live action, lila rose anti-abortion videos*.
https://www.buzzfeednews.com/article/claudiakoerner/facebook-fact-check-abortion-video-doctors-medical, January 2022.
(Accessed on 01/05/2022).

[22] *Gephi - the open graph viz platform*.
https://gephi.org/, Jan 2022.
(Accessed on 01/08/2022).

[23] *Google fact check feature: What it means for your online efforts - act-on*.

https://act-on.com/blog/google-fact-check-feature-what-it-means-for-your-online-efforts/, Jan 2022.
(Accessed on 01/05/2022).

[24] *Hci and the u.s. presidential election: A few thoughts on a research agenda | by brent hecht | medium*.
https://brenthecht.medium.com/hci-and-the-u-s-presidential-election-a-few-thoughts-on-a-research-agenda-7c1a0a04986, January 2022.
(Accessed on 01/05/2022).

[25] *Industry-leading vector graphics software | adobe illustrator*.
https://www.adobe.com/products/illustrator.html, January 2022.
(Accessed on 01/07/2022).

[26] *Kapwing: The collaborative online video editor*.
https://www.kapwing.com/, January 2022.
(Accessed on 01/07/2022).

[27] *Mooc.org | massive open online courses | an edx site*.
https://www.mooc.org/, January 2022.
(Accessed on 01/08/2022).

[28] *Official adobe photoshop | photo and design software*.
https://www.adobe.com/products/photoshop.html, January 2022.
(Accessed on 01/07/2022).

[29] *See fact checks in youtube search results - youtube help*.
https://support.google.com/youtube/answer/9229632?hl=en, Jan 2022.
(Accessed on 01/05/2022).

[30] *Dataleads*.
https://dataleads.co.in/, 2023.
(Accessed on 06/25/2023).

[31] *Dpa german press agency*.
https://www.dpa.com/en, 2023.
(Accessed on 06/25/2023).

[32] *Triggering google suggests - fatrank.*
https://www.fatrank.com/triggering-google-suggests/, June 2023.
(Accessed on 06/05/2023).

[33] *The african network of centers for investigativereporting's investigative lab*, accessed in 2021.
https://investigate.africa/.

[34] 10UNDER100, *20 eye opening amazon statistics & facts for 2020*, (2020).
https://10under100.com/amazon-statistics-facts/.

[35] A. ABDUL, J. VERMEULEN, D. WANG, B. Y. LIM, AND M. KANKANHALLI, *Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda*, in Proceedings of the 2018 CHI conference on human factors in computing systems, 2018, pp. 1–18.

[36] A. ABDUL, J. VERMEULEN, D. WANG, B. Y. LIM, AND M. KANKANHALLI, *Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda*, Association for Computing Machinery, New York, NY, USA, 2018, p. 1–18.

[37] A. ABILOV, Y. HUA, H. MATATOV, O. AMIR, AND M. NAAMAN, *Voterfraud2020: a multi-modal dataset of election fraud claims on twitter*, Proceedings of the International AAAI Conference on Web and Social Media, 15 (2021), pp. 901–912.

[38] J. ABRAHAM AND M. C. REDDY, *Re-coordinating activities: an investigation of articulation work in patient transfers*, in Proceedings of the 2013 conference on Computer supported cooperative work, 2013, pp. 67–78.

[39] M. S. ACKERMAN, *The intellectual challenge of cscw: The gap between social requirements and technical feasibility*, Human–Computer Interaction, 15 (2000), pp. 179–203.

[40] A. ADADI AND M. BERRADA, *Peeking inside the black-box: A survey on explainable artificial intelligence (xai)*, IEEE Access, 6 (2018), pp. 52138–52160.

[41] B. ADAIR, *The future of fact-checking is all about structured data*, April 2021.

[42] K. ADNAN, R. AKBAR, AND K. S. WANG, *Information extraction from multi-faceted unstructured big data*, International Journal of Recent Technology and Engineering (IJRTE), 8 (2019), pp. 1398–1404.

[43] AFP, *Fact check |*.
`https://factcheck.afp.com//`, April 2021.
(Accessed on 04/15/2021).

[44] A. AGADJANIAN, N. BAKHRU, V. CHI, D. GREENBERG, B. HOLLANDER, A. HURT, J. KIND, R. LU, A. MA, B. NYHAN, ET AL., *Counting the pinocchios: The effect of summary fact-checking data on perceived accuracy and favorability of politicians*, Research & Politics, 6 (2019), p. 2053168019870351.

[45] J. ALBRIGHT, *Untrue tube – youtube's conspiracy ecosystem*, 2018.

[46] H. ALLCOTT AND M. GENTZKOW, *Social media and fake news in the 2016 election*, Journal of economic perspectives, 31 (2017), pp. 211–36.

[47] J. ALMEIDA, *Misinformation dissemination on the web*, in Companion Proceedings of the 2019 World Wide Web Conference, 2019, pp. 740–740.

[48] M. ALRUBAIAN, M. AL-QURISHI, M. M. HASSAN, AND A. ALAMRI, *A credibility analysis system for assessing information on twitter*, IEEE Transactions on Dependable and Secure Computing, 15 (2016), pp. 661–674.

[49] M. A. AMAZEEN, *A critical assessment of fact-checking in 2012*, 2013.

[50] M. A. AMAZEEN, C. J. VARGO, AND T. HOPP, *Reinforcing attitudes in a gatewatching news era: Individual-level antecedents to sharing fact-checks on social media*, Communication Monographs, 86 (2019), pp. 112–132.

[51] I. ARCHIVE, *Internet archive: Wayback machine*.
`https://archive.org/web/`, April 2021.
(Accessed on 04/15/2021).

[52] ARCHIVE.TODAY, *archive.today - wikipedia*.
`https://en.wikipedia.org/wiki/Archive.today`, April 2021.
(Accessed on 04/15/2021).

[53] R. L. ARMSTRONG, *New survey suggests 10% of americans believe the moon landing was fake*, (2019).

[54] A. ARSHT AND D. ETCOVITCH, *The human cost of online content moderation*, Harvard Law Review Online, Harvard University, Cambridge, MA, USA. Retrieved from https://jolt. law. harvard. edu/digest/the-human-cost-ofonline-content-moderation, (2018).

[55] V. ARYA, R. K. BELLAMY, P.-Y. CHEN, A. DHURANDHAR, M. HIND, S. C. HOFFMAN, S. HOUDE, Q. V. LIAO, R. LUSS, A. MOJSILOVIĆ, ET AL., *One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques*, arXiv preprint arXiv:1909.03012, (2019).

[56] M. ASHOORI AND J. D. WEISZ, *In ai we trust? factors that influence trustworthiness of ai-infused decision-making processes*, arXiv preprint arXiv:1912.02675, (2019).

[57] P. BALL AND A. MAXMEN, *The epic battle against coronavirus misinformation and conspiracy theories*, (2020).

[58] BALLOTPEDIA, *The methodologies of fact-checking*, accessed in March 2021.

[59] J. BANDY, *Problematic machine behavior: A systematic literature review of algorithm audits*, Proc. ACM Hum.-Comput. Interact., 5 (2021).

[60] S. BARBOSA AND S. MILAN, *Do not harm in private chat apps: Ethical issues for research on and with whatsapp*, Westminster Papers in Communication and Culture, 14 (2019), pp. 49–65.

[61] M. BASOL, J. ROOZENBEEK, AND S. VAN DER LINDEN, *Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news*, Journal of cognition, 3 (2020).

[62] BBC NEWS, *Measles: Four european nations lose eradication status*, (2019).

[63] R. T. BECKWITH, *Us primaries: Election deniers go door-to-door to confront voters after losses - bloomberg*. `https://www.bloomberg.com/news/articles/2022-08-23/election-deniers-go-door-to-door-to-confront-voters-after-losses?leadSource=uverify%20wall`, 2022. (Accessed on 09/07/2022).

[64] W. BELLAMY, *Malaysia airlines flight 370 final report inconclusive*, (2019).

[65] J. BELLUZ, *Amazon is a giant purveyor of medical quackery*, (2016).

[66] A. BERMES, *Information overload and fake news sharing: A transactional stress perspective exploring the mitigating role of consumers' resilience during covid-19*, Journal of Retailing and Consumer Services, 61 (2021), p. 102555.

[67] J. C. BERTOT AND H. CHOI, *Big data and e-government: Issues, policies, and recommendations*, in Proceedings of the 14th Annual International Conference on Digital Government Research, dg.o '13, New York, NY, USA, 2013, Association for Computing Machinery, p. 1–10.

[68] A. BESSI, M. COLETTO, G. A. DAVIDESCU, A. SCALA, G. CALDARELLI, AND W. QUATTROCIOCCHI, *Science vs conspiracy: Collective narratives in the age of misinformation*, PloS one, 10 (2015), p. e0118093.

[69] M. M. BHUIYAN, C. A. BAUTISTA ISAZA, T. MITRA, AND S. W. LEE, *Othertube: Facilitating content discovery and reflection by exchanging youtube recommendations with strangers*, in CHI Conference on Human Factors in Computing Systems, 2022, pp. 1–17.

[70] M. M. BHUIYAN, M. HORNING, S. W. LEE, AND T. MITRA, *Nudgecred: Supporting news credibility assessment on social media through nudges*, Proceedings of the ACM on Human-Computer Interaction, 5 (2021), pp. 1–30.

[71] J. BISBEE, M. BROWN, A. LAI, R. BONNEAU, J. NAGLER, AND J. A. TUCKER, *Election fraud, youtube, and public perception of the legitimacy of president biden*, Journal of Online Trust and Safety, 1 (2022).

[72] A. BONDIELLI AND F. MARCELLONI, *A survey on fake news and rumour detection techniques*, Information Sciences, 497 (2019), pp. 38–55.

[73] N. L. BRAGAZZI, I. BARBERIS, R. ROSSELLI, V. GIANFREDI, D. NUCCI, M. MORETTI, T. SALVATORI, G. MARTUCCI, AND M. MARTINI, *How often people google for vaccination: Qualitative and quantitative insights from a systematic search of the web-based activities using google trends*, Human Vaccines & Immunotherapeutics, 13 (2017), pp. 464–469.
PMID: 27983896.

[74] V. BRAUN AND V. CLARKE, *Using thematic analysis in psychology*, Qualitative research in psychology, 3 (2006), pp. 77–101.

[75] A. BRUNS, *Are filter bubbles real?*, (2019).

[76] A. BRUNS, *Filter bubble*, Internet Policy Review, 8 (2019).

[77] P. BUMP, *The unique role of fox news in the misinformation universe - the washington post*.
https://www.washingtonpost.com/politics/2021/11/08/unique-role-fox-news-misinformation-universe/, 2021.
(Accessed on 09/10/2022).

[78] T. D. BURGESS II AND S. M. SALES, *Attitudinal effects of "mere exposure": A reevaluation*, Journal of Experimental Social Psychology, 7 (1971), pp. 461–472.

[79] J. BURRELL, *How the machine 'thinks': Understanding opacity in machine learning algorithms*, Big Data & Society, 3 (2016), p. 2053951715622512.

[80] BUZZSUMO, *Buzzsumo.com*.
https://buzzsumo.com/, April 2021.
(Accessed on 04/15/2021).

[81] J. C. DOS SANTOS, S. WM SIQUEIRA, B. PEREIRA NUNES, P. P. BALESTRASSI, AND F. RS PEREIRA, *Is there personalization in twitter search? a study on polarized opinions about the brazilian welfare reform*, in 12th ACM Conference on Web Science, 2020, pp. 267–276.

[82] N. CARNE, *'conspiracies' dominate youtube climate modification videos*, (2019).

[83] N. CASS, T. SCHWANEN, AND E. SHOVE, *Infrastructures, intersections and societal transformations*, Technological Forecasting and Social Change, 137 (2018), pp. 160–167.

[84] C. CASTILLO, M. MENDOZA, AND B. POBLETE, *Information credibility on twitter*, in Proceedings of the 20th international conference on World wide web, 2011, pp. 675–684.

[85] M. CAULFIELD, *Web literacy for student fact-checkers*, 2017.

[86] S. CAZALENS, P. LAMARRE, J. LEBLAY, I. MANOLESCU, AND X. TANNIER, *A content management perspective on fact-checking*, in Companion Proceedings of the The Web Conference 2018, 2018, pp. 565–574.

[87] A. S. CENTRAL, *Dietary supplements*, (accessed in 2020).

[88] A. CERONE, E. NAGHIZADE, F. SCHOLER, D. MALLAL, R. SKELTON, AND D. SPINA, *Watch'n'check: Towards a social media monitoring tool to assist fact-checking experts*, in 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2020, pp. 607–613.

[89] B. CHAUDHURI, *Paradoxes of intermediation in aadhaar: Human making of a digital infrastructure*, South Asia: Journal of South Asian Studies, 42 (2019), pp. 572–587.

[90] N. V. CHAWLA, K. W. BOWYER, L. O. HALL, AND W. P. KEGELMEYER, *Smote: synthetic minority over-sampling technique*, Journal of artificial intelligence research, 16 (2002), pp. 321–357.

[91] A. CHECK, *Africa check | sorting fact from fiction*. `https://africacheck.org/`, April 2021. (Accessed on 04/15/2021).

[92] J. M. CHEJFEC-CIOCIANO, J. P. MARTÍNEZ-HERRERA, A. D. PARRA-GUERRA, R. CHEJFEC, F. J. BARBOSA-CAMACHO, J. C. IBARROLA-PEÑA, G. CERVANTES-GUEVARA, G. A. CERVANTES-CARDONA, C. FUENTES-OROZCO, E. CERVANTES-PÉREZ, ET AL., *Misinformation about and interest in chlorine dioxide during the covid-19 pandemic in mexico identified using google trends data: infodemiology study*, JMIR infodemiology, 2 (2022), p. e29894.

[93] L. CHEN, R. MA, A. HANNÁK, AND C. WILSON, *Investigating the impact of gender on rank in resume search engines*, in Proceedings of the 2018 chi conference on human factors in computing systems, 2018, pp. 1–14.

[94] L. CHEN, A. MISLOVE, AND C. WILSON, *Peeking beneath the hood of uber*, in Proceedings of the 2015 internet measurement conference, 2015, pp. 495–508.

[95] L. CHEN, A. MISLOVE, AND C. WILSON, *An empirical analysis of algorithmic pricing on amazon marketplace*, in Proceedings of the 25th International Conference on World Wide Web, 2016, pp. 1339–1349.

[96] Q. CHEN, Y. ZHANG, R. EVANS, AND C. MIN, *Why do citizens share covid-19 fact-checks posted by chinese government social media accounts? the elaboration likelihood model*, International Journal of Environmental Research and Public Health, 18 (2021), p. 10058.

[97] X. CHEN, S.-C. J. SIN, Y.-L. THENG, AND C. S. LEE, *Deterring the spread of misinformation on social network sites: A social cognitive theory-guided intervention*, Proceedings of the Association for Information Science and Technology, 52 (2015), pp. 1–4.

[98] X. CHEN, S.-C. J. SIN, Y.-L. THENG, AND C. S. LEE, *Why students share misinformation on social media: Motivation, gender, and study-level differences*, The Journal of Academic Librarianship, 41 (2015), pp. 583–592.

[99] D. CHERUIYOT AND R. FERRER-CONILL, *"fact-checking africa"*, Digital Journalism, 6 (2018), pp. 964–975.

[100] P. CHONKA, S. DIEPEVEEN, AND Y. HAILE, *Algorithmic power and african indigenous languages: search engine autocomplete and the global multilingual internet*, Media, Culture & Society, 45 (2023), pp. 246–265.

[101] CISCO, *Cisco visual networking index: Forecast and trends, 2017–2022 white paper*, (2019).

[102] J. CONDITT, *Google partners with fact-checking network to fight fake news*, 2017.

[103] W. CONTRIBUTORS, *Occam's razor — Wikipedia, the free encyclopedia*, 2022. [Online; accessed 13-September-2022].

[104] J. COOK, S. LEWANDOWSKY, AND U. K. ECKER, *Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence*, PloS one, 12 (2017), p. e0175799.

[105] A. COSSARD, G. D. F. MORALES, K. KALIMERI, Y. MEJOVA, D. PAOLOTTI, AND M. STARNINI, *Falling into the echo chamber: the italian vaccination debate on twitter*, in Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, 2020, pp. 130–140.

[106] P. COVINGTON, J. ADAMS, AND E. SARGIN, *Deep neural networks for youtube recommendations*, in Proceedings of the 10th ACM Conference on Recommender Systems, 2016.

[107] CROWDTANLE, *Crowdtangle | content discovery and social monitoring made easy*. `https://www.crowdtangle.com/`, April 2021. (Accessed on 04/15/2021).

[108] P. M. DAHLGREN, *A critical review of filter bubbles and a comparison with selective exposure*, Nordicom Review, 42 (2021), pp. 15–33.

[109] E. DAI, Y. SUN, AND S. WANG, *Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository*, Proceedings of the International AAAI Conference on Web and Social Media, 14 (2020), pp. 853–862.

[110] F. J. DAMERAU, *A technique for computer detection and correction of spelling errors*, Communications of the ACM, 7 (1964), pp. 171–176.

[111] M. DANILEVSKY, K. QIAN, R. AHARONOV, Y. KATSIS, B. KAWAS, AND P. SEN, *A survey of the state of explainable AI for natural language processing*, in Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, Suzhou, China, Dec. 2020, Association for Computational Linguistics, pp. 447–459.

[112] E. DAYTON, *Amazon statistics you should know: Opportunities to make the most of america's top online marketplace*.

[113] S. DE PAOR AND B. HERAVI, *Information literacy and fake news: How the field of librarianship can help combat the epidemic of fake news*, The Journal of Academic Librarianship, 46 (2020), p. 102218.

[114] B. DEAN, *Here's what we learned about organic click through rate*, (2019).

[115] K. C. DESOUZA AND K. L. SMITH, *Big data for social innovation*, Stanford Social Innovation Review, 12 (2014), pp. 38–43.

[116] J. DEVLIN, M.-W. CHANG, K. LEE, AND K. TOUTANOVA, *Bert: Pre-training of deep bidirectional transformers for language understanding*, arXiv preprint arXiv:1810.04805, (2018).

[117] N. DIAKOPOULOS, D. TRIELLI, J. STARK, AND S. MUSSENDEN, *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).

[118] N. DIAS AND A. SIPPITT, *Researching fact checking: Present limitations and future opportunities*, The Political Quarterly, 91 (2020), pp. 605–613.

[119] C. DICKEY, *The rise and fall of facts*, 2019.

[120] R. DIRESTA, *The complexity of simply searching for medical advice*, 2018.

[121] R. DIRESTA, *How amazon's algorithms curated a dystopian bookstore*, (2019). https://www.wired.com/story/amazon-and-the-spread-of-health-misinformation/.

[122] J. D'ONFRO, *Youtube adding wikipedia links debunking conspiracy theories*. https://www.cnbc.com/2018/03/13/youtube-wikipedia-links-debunk-conspiracy.html, 2018.
(Accessed on 09/12/2022).

[123] B. DOSONO AND B. SEMAAN, *Moderation practices as emotional labor in sustaining online communities: The case of aapi identity work on reddit*, in Proceedings of the 2019 CHI conference on human factors in computing systems, 2019, pp. 1–13.

[124] K. M. DOUGLAS, R. M. SUTTON, D. JOLLEY, AND M. J. WOOD, *The social, political, environmental, and health-related consequences of conspiracy theories*, The psychology of conspiracy, (2015), pp. 183–200.

[125] DPA, *Fact check*. https://dps-factify.com, Aug 2021.

[126] F. DRAFT, *Combating misinformation in under-resourced languages: lessons from around the world*, 2020.

[127] T. DREISBACH, *On amazon, dubious 'antiviral' supplements proliferate amid pandemic*, (2020).

[128] A. F. DUGAS, Y.-H. HSIEH, S. R. LEVIN, J. M. PINES, D. P. MAREINISS, A. MOHAREB, C. A. GAYDOS, T. M. PERL, AND R. E. ROTHMAN, *Google flu trends: correlation with emergency department influenza rates and crowding metrics*, Clinical infectious diseases, 54 (2012), pp. 463–469.

[129] C. DWYER, *Task technology fit, the social technical gap and social networking sites*, AMCIS 2007 Proceedings, (2007), p. 374.

[130] M. DYE, D. NEMER, N. KUMAR, AND A. S. BRUCKMAN, *If it rains, ask grandma to disconnect the nano: Maintenance & care in havana's streetnet*, Proceedings of the ACM on human-computer interaction, 3 (2019), pp. 1–27.

[131] M. DYE, D. NEMER, J. MANGIAMELI, A. S. BRUCKMAN, AND N. KUMAR, *El paquete semanal: The week's internet in havana*, in Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018, pp. 1–12.

[132] U. K. ECKER, Z. O'REILLY, J. S. REID, AND E. P. CHANG, *The effectiveness of short-format refutational fact-checks*, British Journal of Psychology, 111 (2020), pp. 36–54.

[133] J. ELIZABETH, *Who are you calling a fact checker?*, 2014.

[134] R. EPSTEIN AND R. E. ROBERTSON, *The search engine manipulation effect (seme) and its possible impact on the outcomes of elections*, Proceedings of the National Academy of Sciences, 112 (2015), pp. E4512–E4521.

[135] R. EPSTEIN, R. E. ROBERTSON, D. LAZER, AND C. WILSON, *Suppressing the search engine manipulation effect (seme)*, Proceedings of the ACM on Human-Computer Interaction, 1 (2017), pp. 1–22.

[136] S. ERDEN, K. NALBANT, AND H. FERAHKAYA, *Autism and vaccinations: Does google side with science?*, Journal of Contemporary Medicine, 9 (2019), pp. 295–299.

[137] I. ETIKAN, S. A. MUSA, AND R. S. ALKASSIM, *Comparison of convenience sampling and purposive sampling*, American journal of theoretical and applied statistics, 5 (2016), pp. 1–4.

[138] F. FACT, *coof-2020.pdf*. `https://fullfact.org/media/uploads/coof-2020.pdf`, Dec 2020. (Accessed on 07/27/2023).

[139] ——, *Full fact*. `https://fullfact.org/`, April 2021. (Accessed on 04/15/2021).

[140] M. FADDOUL, G. CHASLOT, AND H. FARID, *A longitudinal analysis of youtube's promotion of conspiracy videos*, arXiv preprint arXiv:2003.03318, (2020).

[141] K. P. FERGUSON, *Impact of technology on rural appalachian health care providers: Assessment of technological infrastructure, behaviors, and attitudes.*, (2005).

[142] J. FERREIRA, H. SHARP, AND H. ROBINSON, *User experience design and agile development: managing cooperation through articulation work*, Software: Practice and Experience, 41 (2011), pp. 963–974.

[143] T. FINANCIAL, *Amazon removed 1 million fake coronavirus cures and overpriced products*, (2020).

[144] FIRST DRAFT, *The importance of local context in taking on misinformation: Lessons from africacheck*, 2020.

[145] S. FISCHER, *Amazon has a misinformation problem, too*, (2019).

[146] R. S. FISH, R. E. KRAUT, AND M. D. LELAND, *Quilt: A collaborative tool for cooperative writing*, in Proceedings of the ACM SIGOIS and IEEECS TC-OA 1988 conference on Office information systems, 1988, pp. 30–37.

[147] C. FOR AN INFORMED PUBLIC, D. F. R. LAB, GRAPHIKA, AND S. I. OBSERVATORY, *The long fuse: Misinformation and the 2020 election*, (2021).

[148] S. FOX, *Online health search 2006*, (2006).

[149] K. FRIDKIN, P. J. KENNEY, AND A. WINTERSIECK, *Liar, liar, pants on fire: How fact-checking influences citizens' reactions to negative advertising*, Political Communication, 32 (2015), pp. 127–151.

[150] A. FRIGGERI, L. ADAMIC, D. ECKLES, AND J. CHENG, *Rumor cascades*, in Eighth International AAAI Conference on Weblogs and Social Media, 2014.

[151] A. J. B. FROM, *Communicating fact checks online*, (2020).

[152] C. M. FULLER, D. P. BIROS, AND R. L. WILSON, *Decision support for determining veracity via linguistic-based cues*, Decision Support Systems, 46 (2009), pp. 695–703.

[153] FULLFACT, *Github - fullfact/claim-review-schema-wordpress-plugin: An open source project to create a wordpress plugin for claim review schema.* `https://github.com/FullFact/claim-review-schema-wordpress-plugin`, April 2021.
(Accessed on 04/15/2021).

[154] E. GAILLARD, *Facebook under fire for permitting anti-vax groups*, (2019).

[155] H. GAO, X. WANG, G. BARBIER, AND H. LIU, *Promoting coordination for disaster relief–from crowdsourcing to coordination*, in International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, Springer, 2011, pp. 197–204.

[156] R. K. GARRETT AND B. E. WEEKS, *The promise and peril of real-time corrections to political misperceptions*, in Proceedings of the 2013 conference on Computer supported cooperative work, 2013, pp. 1047–1058.

[157] Y. GE, S. LIU, R. GAO, Y. XIAN, Y. LI, X. ZHAO, C. PEI, F. SUN, J. GE, W. OU, ET AL., *Towards long-term fairness in recommendation*, in Proceedings of the 14th ACM International Conference on Web Search and Data Mining, 2021, pp. 445–453.

[158] A. GHENAI AND Y. MEJOVA, *Catching zika fever: Application of crowdsourcing and machine learning for tracking health misinformation on twitter*, arXiv preprint arXiv:1707.03778, (2017).

[159] A. GHENAI AND Y. MEJOVA, *Fake cures: user-centric modeling of health misinformation in social media*, Proceedings of the ACM on human-computer interaction, 2 (2018), pp. 1–20.

[160] P. GHEZZI, P. G. BANNISTER, G. CASINO, A. CATALANI, M. GOLDMAN, J. MORLEY, M. NEUNEZ, A. PRADOS-BO, P. R. SMEESTERS, M. TADDEO, ET AL., *Online information of vaccines: information quality, not only privacy, is an ethical responsibility of search engines*, Frontiers in Medicine, 7 (2020).

[161] T. GILLESPIE, *The relevance of algorithms*, Media technologies: Essays on communication, materiality, and society, 167 (2014).

[162] T. GILLESPIE, *Algorithmically recognizable: Santorum's google problem, and google's santorum problem*, Information, communication & society, 20 (2017), pp. 63–80.

[163] F. GIRARDIN, *Towards Reducing the Social-Technical Gap in Location-Aware Computing*, PhD thesis, Citeseer, 2007.

[164] A. GLASER, *Amazon is suggesting "frequently bought together" items that can make a bomb*, (2017).

[165] O. GOLDHILL, *Amazon is selling coronavirus misinformation*, (2020).

[166] A. GOLDMAN AND C. O'CONNOR, *Social Epistemology*, in The Stanford Encyclopedia of Philosophy, E. N. Zalta, ed., Metaphysics Research Lab, Stanford University, Winter 2021 ed., 2021.

[167] M. GOLEBIEWSKI AND D. BOYD, *Data voids: Where missing data can easily be exploited*, (2019).

[168] L. A. GOODMAN, *Snowball sampling*, The annals of mathematical statistics, (1961), pp. 148–170.

[169] GOOGLE, *Google's search quality rating guidelines*, (2019).

[170] L. GRACE AND B. HONE, *Factitious: Large scale computer game to fight fake news and improve news literacy*, in Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–8.

[171] D. GRAVES, *Understanding the promise and limits of automated fact-checking*, (2018).

[172] L. GRAVES, *Deciding what's true: Fact-checking journalism and the new ecology of news*, PhD thesis, Columbia University, 2013.

[173] L. GRAVES, *Anatomy of a fact check: Objective practice and the contested epistemology of fact checking*, Communication, Culture & Critique, 10 (2017), pp. 518–537.

[174] L. GRAVES AND M. A. AMAZEEN, *Fact-checking as idea and practice in journalism*, in Oxford Research Encyclopedia of Communication, 2019.

[175] L. GRAVES AND C. W. ANDERSON, *Discipline and promote: Building infrastructure and managing algorithms in a "structured journalism" project by professional fact-checking groups*, New Media & Society, 22 (2020), pp. 342–360.

[176] L. GRAVES AND F. CHERUBINI, *The rise of fact-checking sites in europe*, (2016).

[177] L. GRAVES AND T. GLAISYER, *The fact-checking universe in spring 2012*, New America, (2012).

[178] J. GREEN, W. HOBBS, S. MCCABE, AND D. LAZER, *Online engagement with 2020 election misinformation and turnout in the 2021 georgia runoff election*, Proceedings of the National Academy of Sciences, 119 (2022), p. e2115900119.

[179] R. E. GRINTER, *Using a configuration management tool to coordinate software development*, in Proceedings of conference on Organizational computing systems, 1995, pp. 168–177.

[180] T. GRØNSUND AND M. AANESTAD, *Augmenting the algorithm: Emerging human-in-the-loop work configurations*, The Journal of Strategic Information Systems, 29 (2020), p. 101614.
Strategic Perspectives on Digital Work and Organizational Transformation.

[181] Z. GUAN AND E. CUTRELL, *An eye tracking study of the effect of target rank on web search*, in Proceedings of the SIGCHI conference on Human factors in computing systems, 2007, pp. 417–420.

[182] F. GUERRA, D. LINZ, R. GARCIA, B. KOMMATA, J. KOSIUK, J. CHUN, S. BOVEDA, AND D. DUNCKER, *The use of instant messaging in clinical data sharing: the ehra sms survey*, EP Europace, 23 (2021), pp. euab116–515.

[183] A. GUESS, J. NAGLER, AND J. TUCKER, *Less than you think: Prevalence and predictors of fake news dissemination on facebook*, Science advances, 5 (2019), p. eaau4586.

[184] A. GUPTA, H. LAMBA, P. KUMARAGURU, AND A. JOSHI, *Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy*, in Proceedings of the 22nd international conference on World Wide Web, ACM, 2013, pp. 729–736.

[185] P. H. CARSTENSEN, *Modeling coordination work: Lessons learned from analyzing a cooperative work setting*, in Symbiosis of Human and Artifact, Y. Anzai, K. Ogawa, and H. Mori, eds., vol. 20 of Advances in Human Factors/Ergonomics, Elsevier, 1995, pp. 327–332.

[186] J. HALE, *More than 500 hours of content are now being uploaded to youtube every minute*, (2019).

[187] A. HANNAK, P. SAPIEZYNSKI, A. MOLAVI KAKHKI, B. KRISHNAMURTHY, D. LAZER, A. MISLOVE, AND C. WILSON, *Measuring personalization of web search*, in Proceedings of the 22nd international conference on World Wide Web, 2013, pp. 527–538.

[188] A. HANNAK, G. SOELLER, D. LAZER, A. MISLOVE, AND C. WILSON, *Measuring price discrimination and steering on e-commerce web sites*, in Proceedings of the 2014 conference on internet measurement conference, 2014, pp. 305–318.

[189] A. HANNÁK, C. WAGNER, D. GARCIA, A. MISLOVE, M. STROHMAIER, AND C. WILSON, *Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr*, in Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW '17, ACM, 2017, pp. 1914–1933.

[190] M. M. HAQUE, M. YOUSUF, A. S. ALAM, P. SAHA, S. I. AHMED, AND N. HASSAN, *Combating misinformation in bangladesh: Roles and responsibilities as perceived by journalists, fact-checkers, and users*, Proceedings of the ACM on Human-Computer Interaction, 4 (2020), pp. 1–32.

[191] N. HASSAN, F. ARSLAN, C. LI, AND M. TREMAYNE, *Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster*, in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 1803–1812.

[192] N. HASSAN, C. LI, AND M. TREMAYNE, *Detecting check-worthy factual claims in presidential debates*, in Proceedings of the 24th acm international on conference on information and knowledge management, 2015, pp. 1835–1838.

[193] N. HASSAN, M. YOUSUF, M. MAHFUZUL HAQUE, J. A. SUAREZ RIVAS, AND M. KHADIMUL ISLAM, *Examining the roles of automation, crowds and professionals towards sustainable fact-checking*, in Companion Proceedings of The 2019 World Wide Web Conference, 2019, pp. 1001–1006.

[194] F. HE AND S. HAN, *A method and tool for human–human interaction and instant collaboration in cscw-based cad*, Computers in Industry, 57 (2006), pp. 740–751.

[195] R. HEEKS, *Most egovernment-for-development projects fail: how can risks be reduced?*, (2003).

[196] G. T. HELP, *Explore results by region*, (2020).

[197] M. HERR, *Writing and Publishing Your Book: A Guide for Experts in Every Field*, ABC-CLIO, 2017.

[198] B. D. HORNE AND S. ADALI, *This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news*, in Eleventh International AAAI Conference on Web and Social Media, 2017.

[199] R. HOU, V. PÉREZ-ROSAS, S. LOEB, AND R. MIHALCEA, *Towards automatic detection of misinformation in online medical videos*, in 2019 International conference on multimodal interaction, 2019, pp. 235–243.

[200] D. HOULI, M. L. RADFORD, AND V. K. SINGH, *"covid19 is_": The perpetuation of coronavirus conspiracy theories via google autocomplete*, Proceedings of the Association for Information Science and Technology, 58 (2021), pp. 218–229.

[201] D. HU, S. JIANG, R. E. ROBERTSON, AND C. WILSON, *Auditing the partisanship of google search snippets*, in The World Wide Web Conference, WWW '19, New York, NY, USA, 2019, Association for Computing Machinery, p. 693–704.

[202] D. HU, S. JIANG, R. E. ROBERTSON, AND C. WILSON, *Auditing the partisanship of google search snippets*, in The World Wide Web Conference, WWW '19, ACM, 2019, pp. 693–704.

[203] E. HUGHES, R. WANG, P. JUNEJA, T. MITRA, AND A. X. ZHANG, *Introducing credibility signals and citations to video-sharing platforms*, (2021).

[204] K. HUNT, P. AGARWAL, AND J. ZHUANG, *Monitoring misinformation on twitter during crisis events: a machine learning approach*, Risk analysis, (2020).

[205] E. HUSSEIN, P. JUNEJA, AND T. MITRA, *Measuring misinformation in video search platforms: An audit study on youtube*, Proceedings of the ACM on Human-Computer Interaction, 4 (2020), pp. 1–27.

[206] A. HUTCHINSONA, *Pinterest will limit search results for vaccine-related queries to content from official health outlets*, (2019).

[207] INFLUENCER, *Find everything about youtube on noxinfluencer*. `https://www.noxinfluencer.com/`, April 2021. (Accessed on 04/15/2021).

[208] S. INSKEEP, *Timeline: The false election fraud story trump told for months before jan. 6 : Npr*.

https://www.npr.org/2021/02/08/965342252/timeline-what-trump-told-supporters-for-months-before-they-attacked, 2022.
(Accessed on 04/18/2022).

[209] INVID, *Invid verification plugin - invid project*.
https://www.invid-project.eu/tools-and-services/invid-verification-plugin/, April 2021.
(Accessed on 04/15/2021).

[210] M. JACK, J. CHEN, AND S. J. JACKSON, *Infrastructure as creative action: Online buying, selling, and delivery in phnom penh*, in Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 2017, pp. 6511–6522.

[211] S. J. JACKSON, *11 rethinking repair*, Media technologies: Essays on communication, materiality, and society, (2014), pp. 221–39.

[212] S. J. JACKSON, A. POMPE, AND G. KRIESHOK, *Repair worlds: maintenance, repair, and ict for development in rural namibia*, in Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, 2012, pp. 107–116.

[213] S. M. JANG AND J. K. KIM, *Third person effects of fake news: Fake news regulation and media literacy interventions*, Computers in human behavior, 80 (2018), pp. 295–302.

[214] A. JAZEERA, *Breaking news, world news and video from al jazeera | today's latest from al jazeera*.
https://www.aljazeera.com/, April 2021.
(Accessed on 04/15/2021).

[215] S. JIANG, S. BAUMGARTNER, A. ITTYCHERIAH, AND C. YU, *Factoring fact-checks: Structured information extraction from fact-checking articles*, in Proceedings of The Web Conference 2020, 2020, pp. 1592–1603.

[216] S. JIANG, R. E. ROBERTSON, AND C. WILSON, *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, 2019, pp. 278–289.

[217] S. JIANG AND C. WILSON, *Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media*, Proceedings of the ACM on Human-Computer Interaction, 2 (2018), pp. 1–23.

[218] D. JIMENEZ AND C. LI, *An empirical study on identifying sentences with salient factual statements*, in 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, 2018, pp. 1–8.

[219] M. JIROTKA, R. PROCTER, T. RODDEN, AND G. C. BOWKER, *Collaboration in e-research*, Computer Supported Cooperative Work (CSCW), 15 (2006), pp. 251–255.

[220] H. M. JOHNSON AND C. M. SEIFERT, *Sources of the continued influence effect: When misinformation in memory affects later inferences.*, Journal of experimental psychology: Learning, memory, and cognition, 20 (1994), p. 1420.

[221] G. JORIS, F. DE GROVE, K. VAN DAMME, AND L. DE MAREZ, *News diversity reconsidered: A systematic literature review unraveling the diversity in conceptualizations*, Journalism Studies, 21 (2020), pp. 1893–1912.

[222] P. JUNEJA, M. M. BHUIYAN, AND T. MITRA, *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, 2023, pp. 1–22.

[223] P. JUNEJA AND T. MITRA, *Auditing e-commerce platforms for algorithmically curated vaccine misinformation*, in Proceedings of the 2021 chi conference on human factors in computing systems, 2021, pp. 1–27.

[224] A. KAPLAN, *Youtube has allowed conspiracy theories about interference with voting machines to go viral | media matters for america*.
`https://www.mediamatters.org/google/youtube-has-allowed-conspiracy-theories-about-interference-voting-machines-go-viral`, 2020.
(Accessed on 09/08/2022).

[225] G. KARADZHOV, P. NAKOV, L. MÀRQUEZ, A. BARRÓN-CEDEÑO, AND I. KOYCHEV, *Fully automated fact checking using external sources*, arXiv preprint arXiv:1710.00341, (2017).

[226] H. KARASTI AND J. BLOMBERG, *Studying infrastructuring ethnographically*, Computer Supported Cooperative Work (CSCW), 27 (2018), pp. 233–265.

[227] A. KARP AND B. PARDO, *Hapteq: A collaborative tool for visually impaired audio producers*, in Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences, 2017, pp. 1–4.

[228] Y. S. KARTAL, B. GUVENEN, AND M. KUTLU, *Too many claims to fact-check: Prioritizing political claims based on check-worthiness*, arXiv preprint arXiv:2004.08166, (2020).

[229] A. KATA, *A postmodern pandora's box: anti-vaccination misinformation on the internet*, Vaccine, 28 (2010), pp. 1709–1716.

[230] A. KAZEMI, K. GARIMELLA, G. K. SHAHI, D. GAFFNEY, AND S. A. HALE, *Tiplines to combat misinformation on encrypted platforms: A case study of the 2019 indian election on whatsapp*, arXiv preprint arXiv:2106.04726, (2021).

[231] A. KERR AND J. D. KELLEHER, *The recruitment of passion and community in the service of capital: Community managers in the digital games industry*, Critical studies in media communication, 32 (2015), pp. 177–192.

[232] A. KHARA, *Iran: Over 700 dead after drinking alcohol to cure coronavirus | coronavirus pandemic news | al jazeera*, April 2020.
(Accessed on 06/21/2023).

[233] J. KIM, B. TABIBIAN, A. OH, B. SCHÖLKOPF, AND M. GOMEZ-RODRIGUEZ, *Leveraging the crowd to detect and reduce the spread of fake news and misinformation*, in Proceedings of the eleventh ACM international conference on web search and data mining, 2018, pp. 324–332.

[234] S. KIM, O. F. YALCIN, S. E. BESTVATER, K. MUNGER, B. L. MONROE, AND B. A. DESMARAIS, *The effects of an informational intervention on attention to anti-vaccination content on youtube*, in Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, 2020, pp. 949–953.

[235] C. KLIMAN-SILVER, A. HANNAK, D. LAZER, C. WILSON, AND A. MISLOVE, *Location, location, location: The impact of geolocation on web search personalization*,

in Proceedings of the 2015 Internet Measurement Conference, IMC '15, ACM, 2015, pp. 121–127.

[236] C. KLIMAN-SILVER, A. HANNAK, D. LAZER, C. WILSON, AND A. MISLOVE, *Location, location, location: The impact of geolocation on web search personalization*, in Proceedings of the 2015 Internet Measurement Conference, ACM, 2015, pp. 121–127.

[237] P. KNIGHT, *Outrageous conspiracy theories: Popular and official responses to 9/11 in germany and the united states*, New German Critique, (2008), pp. 165–193.

[238] S. KNOBLOCH-WESTERWICK, B. K. JOHNSON, N. A. SILVER, AND A. WESTERWICK, *Science exemplars in the eye of the beholder: How exposure to online science information affects attitudes*, Science Communication, 37 (2015), pp. 575–601.

[239] N. KOTONYA AND F. TONI, *Explainable automated fact-checking for public health claims*, arXiv preprint arXiv:2010.09926, (2020).

[240] M. KRÜGER, A. WEIBERT, D. D. C. LEAL, D. RANDALL, AND V. WULF, *It takes more than one hand to clap: On the role of 'care'in maintaining design results.*, in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–14.

[241] J. H. KUKLINSKI, P. J. QUIRK, J. JERIT, D. SCHWIEDER, AND R. F. RICH, *Misinformation and the currency of democratic citizenship*, The Journal of Politics, 62 (2000), pp. 790–816.

[242] J. KULSHRESTHA, M. ESLAMI, J. MESSIAS, M. B. ZAFAR, S. GHOSH, K. P. GUMMADI, AND K. KARAHALIOS, *Quantifying search bias: Investigating sources of bias for political searches in social media*, in Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW '17, New York, NY, USA, 2017, Association for Computing Machinery, p. 417–432.

[243] S. KUMAR, R. WEST, AND J. LESKOVEC, *Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes*, in Proceedings of the 25th international conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2016, pp. 591–602.

[244] A. LAMPINEN, V. BELLOTTI, C. CHESHIRE, AND M. GRAY, *Cscw and thesharing economy: The future of platforms as sites of work collaboration and trust*, in Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion, 2016, pp. 491–497.

[245] W. LANGEWIESCHE, *What really happened to malaysia's missing airplane*, (2019).

[246] B. LATOUR AND S. WOOLGAR, *Laboratory life: The construction of scientific facts*, Princeton University Press, 2013.

[247] J. LAXA, *The consumption of disinformation as a health crisis*, Journal of Public Health, 45 (2023), pp. e161–e161.

[248] J. LAZAR, *Public policy and hci: Making an impact in the future*, Interactions, 22 (2015), p. 69–71.

[249] B. LE, D. SPINA, F. SCHOLER, AND H. CHIA, *A crowdsourcing methodology to measure algorithmic bias in black-box systems: A case study with covid-related searches*, in Proceedings of the Third Workshop on Bias and Social Aspects in Search and Recommendation (Bias@ ECIR 2022), 2022.

[250] D. LEADS, *Data leads*.
`https://dataleads.co.in/`, April 2021.
(Accessed on 04/15/2021).

[251] J. LEBLAY, I. MANOLESCU, AND X. TANNIER, *Computational fact-checking: Problems, state of the art, and perspectives*, in The Web Conference, 2018.

[252] C. P. LEE, P. DOURISH, AND G. MARK, *The human infrastructure of cyberinfrastructure*, in Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work, 2006, pp. 483–492.

[253] S.-P. LEINO, S. LIND, M. POYADE, S. KIVIRANTA, P. MULTANEN, A. REYES-LECUONA, A. MÄKIRANTA, AND A. MUHAMMAD, *Enhanced industrial maintenance work task planning by using virtual engineering tools and haptic user interfaces.*, in HCI (13), 2009, pp. 346–354.

[254] P. LEWIS AND E. MCCORMICK, *How an ex-youtube insider investigated its secret algorithm*, (2018).

[255] M. LEWIS-BECK, A. E. BRYMAN, AND T. F. LIAO, *The Sage encyclopedia of social science research methods*, Sage Publications, 2003.

[256] C. LIMA, *Youtube to remove videos claiming mass fraud changed election results - politico*.
https://www.politico.com/news/2020/12/09/youtube-videos-mass-fraud-election-results-443925, 2020.
(Accessed on 09/08/2022).

[257] C. LIMA AND A. SCHAFFER, *Study finds social media posts about election fraud still prevalent - the washington post*.
https://www.washingtonpost.com/politics/2022/08/09/social-media-posts-about-election-fraud-still-prevalent-study-finds/, 2022.
(Accessed on 09/06/2022).

[258] P. LINARDATOS, V. PAPASTEFANOPOULOS, AND S. KOTSIANTIS, *Explainable ai: A review of machine learning interpretability methods*, Entropy, 23 (2021), p. 18.

[259] C. LIOMA, J. G. SIMONSEN, AND B. LARSEN, *Evaluation measures for relevance and credibility in ranked lists*, in Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, 2017, pp. 91–98.

[260] I. LITERAT, Y. K. CHANG, AND S.-Y. HSU, *Gamifying fake news: Engaging youth in the participatory design of news literacy games*, Convergence, 26 (2020), pp. 503–516.

[261] J. LOCHER, *How to fight election misinformation in 2022 – brewminate: A bold blend of news and ideas*.
https://brewminate.com/how-to-fight-election-misinformation-in-2022/, 2022.
(Accessed on 09/08/2022).

[262] R. LUDOLPH, A. ALLAM, AND P. SCHULZ, *Manipulating google's knowledge graph box to counter biased information processing during an online search on vaccination: Application of a technological debiasing strategy*, Journal of Medical Internet Research, 18 (2016), p. e137.

[263] N. LUNDBERG AND H. TELLIOĞLU, *Understanding complex coordination processes in health care*, Scandinavian Journal of Information Systems, 11 (1999), p. 5.

[264] M. MACDONALD AND M. A. BROWN, *Republican candidates are spreading more fake news than just two years ago - the washington post*.
https://www.washingtonpost.com/politics/2022/08/29/republicans-democrats-misinformation-falsehoods/, 2022.
(Accessed on 09/10/2022).

[265] G. MARK, B. AL-ANI, AND B. SEMAAN, *Repairing human infrastructure in war zones*, Proceedings of ISCRAM, (2009), pp. 10–13.

[266] R. MARKUSSEN, *Politics of intervention in design: Feminist reflections on the scandinavian tradition*, ai & Society, 10 (1996), pp. 127–141.

[267] M. MARS AND R. E. SCOTT, *Whatsapp in clinical practice: A literature*, The Promise of New Technologies in an Age of New Health Challenges, (2016), p. 82.

[268] L. MATSAKIS, *Facebook will crack down on anti-vaccine content*, (2019).

[269] L. MCDONALD AND C. O'DONOVAN, *Youtube continues to promote anti-vax videos as facebook prepares to fight medical misinformation*, (2019).

[270] M. MCGEE, *In quality raters' handbook, google adds higher standards for "your money or your life" websites*, (2013).

[271] MEEDAN, *Meedan*.
https://meedan.com/, April 2021.
(Accessed on 04/15/2021).

[272] P. MELO, J. MESSIAS, G. RESENDE, K. GARIMELLA, J. ALMEIDA, AND F. BENEVENUTO, *Whatsapp monitor: A fact-checking system for whatsapp*, in Proceedings of the International AAAI Conference on Web and Social Media, vol. 13, 2019, pp. 676–677.

[273] P. MELO, J. MESSIAS, G. RESENDE, K. GARIMELLA, J. ALMEIDA, AND F. BENEVENUTO, *Whatsapp monitor: A fact-checking system for whatsapp*, Proceedings of the International AAAI Conference on Web and Social Media, 13 (2019), pp. 676–677.

[274] A. MERELLI, *The average anti-vaxxer is probably not who you think she is*, (2015).

[275] D. METAXA, J. S. PARK, J. A. LANDAY, AND J. HANCOCK, *Search media and elections: A longitudinal investigation of political search results*, Proceedings of the ACM on Human-Computer Interaction, 3 (2019), pp. 1–17.

[276] P. T. METAXAS AND Y. PRUKSACHATKUN, *Manipulation of search engine results during the 2016 us congressional elections*, (2017).

[277] R. MIHAILA, *How to stop a youtube channel from showing up in search results*. `https://www.makeuseof.com/block-youtube-channel-from-search-results`, Feb 2023.
(Accessed on 05/30/2023).

[278] T. MILLER, *Explanation in artificial intelligence: Insights from the social sciences*, Artificial intelligence, 267 (2019), pp. 1–38.

[279] T. MITRA, *Understanding social media credibility*, PhD thesis, Georgia Institute of Technology, 2017.

[280] T. MITRA, S. COUNTS, AND J. W. PENNEBAKER, *Understanding anti-vaccination attitudes in social media*, in Tenth International AAAI Conference on Web and Social Media, 2016.

[281] T. MITRA AND E. GILBERT, *Credbank: A large-scale social media corpus with associated credibility annotations*, in Ninth International AAAI Conference on Web and Social Media, 2015.

[282] T. MITRA, C. J. HUTTO, AND E. GILBERT, *Comparing person-and process-centric strategies for obtaining quality data on amazon mechanical turk*, in Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 2015, pp. 1345–1354.

[283] T. MITRA, G. P. WRIGHT, AND E. GILBERT, *A parsimonious language model of social media credibility across disparate events*, in Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing, 2017, pp. 126–145.

[284] B. MØNSTED AND S. LEHMANN, *Algorithmic detection and analysis of vaccine-denialist sentiment clusters in social networks*, arXiv preprint arXiv:1905.12908, (2019).

[285] J. T. MORGAN, M. GILBERT, D. W. MCDONALD, AND M. ZACHRY, *Project talk: Coordination work and group membership in wikiprojects*, in Proceedings of the 9th International Symposium on Open Collaboration, 2013, pp. 1–10.

[286] D. MOTA, C. V. DE CARVALHO, AND L. P. REIS, *Fostering collaborative work between educators in higher education*, in 2011 IEEE International Conference on Systems, Man, and Cybernetics, IEEE, 2011, pp. 1286–1291.

[287] P. NAKOV, D. CORNEY, M. HASANAIN, F. ALAM, T. ELSAYED, A. BARRÓN-CEDEÑO, P. PAPOTTI, S. SHAAR, AND G. D. S. MARTINO, *Automated fact-checking for assisting human fact-checkers*, arXiv preprint arXiv:2103.07769, (2021).

[288] P. M. NAPOLI, *Exposure diversity reconsidered*, Journal of information policy, 1 (2011), pp. 246–259.

[289] B. A. NARDI AND Y. ENGESTRÖM, *A web on the wind: The structure of invisible work*, Computer supported cooperative work, 8 (1999), pp. 1–8.

[290] NASA, *Nasa facts*, (2001).

[291] B. NEWS, *9/11 conspiracy theories: How they've evolved*, (2011).

[292] A. T. NGUYEN, A. KHAROSEKAR, S. KRISHNAN, S. KRISHNAN, E. TATE, B. C. WALLACE, AND M. LEASE, *Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking*, in Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology, 2018, pp. 189–199.

[293] L. U. NGUYEN, *Infrastructural action in vietnam: Inverting the techno-politics of hacking in the global south*, New Media & Society, 18 (2016), pp. 637–652.

[294] N. OCEANIC AND A. ADMINISTRATION, *Do contrails affect conditions on the surface?*, (2016).

[295] A. OELDORF-HIRSCH, M. SCHMIERBACH, A. APPELMAN, AND M. P. BOYLE, *The ineffectiveness of fact-checking labels on news memes and articles*, Mass Communication and Society, 23 (2020), pp. 682–704.

[296] H. OF COMMONS, *Public administration and constitutional affairs committee, oral evidence: Governance of statistics*, 2019.

[297] A. OLSHANSKY, *Conspiracy theorizing and religious motivated reasoning: Why the earth 'must' be flat*, (2018).

[298] W. H. ORGANIZATION, *Mmr and autism*, (2019).

[299] ——, *Ten threats to global health in 2019*, 2019.

[300] L. H. OWEN, *Republicans seem more susceptible to fake news than democrats (but liberals, don't feel too comfy yet) | nieman journalism lab*. https://www.niemanlab.org/2017/05/republicans-seem-more-susceptible-to-fake-news-than-democrats-but-liberals-dont-feel-too-comfy-yet/, 2017. (Accessed on 09/14/2022).

[301] L. H. OWEN, *One group that's really benefited from covid-19: Anti-vaxxers*, (2020).

[302] A. PÁEZ, *The pragmatic turn in explainable artificial intelligence (xai)*, Minds and Machines, 29 (2019), pp. 441–459.

[303] K. PAPADAMOU, S. ZANNETTOU, J. BLACKBURN, E. DE CRISTOFARO, G. STRINGHINI, AND M. SIRIVIANOS, *"it is just a flu": Assessing the effect of watch history on youtube's pseudoscientific video recommendations*, in Proceedings of the International AAAI Conference on Web and Social Media, vol. 16, 2022, pp. 723–734.

[304] E. PARISER, *The filter bubble: How the new personalized web is changing what we read and how we think*, Penguin, 2011.

[305] F. PASQUALE, *Beyond innovation and competition: The need for qualified transparency in internet intermediaries*, Nw. UL Rev., 104 (2010), p. 105.

[306] F. PASQUALE, *Restoring transparency to automated authority*, J. on Telecomm. & High Tech. L., 9 (2011), p. 235.

[307] S. R. PENDSE, F. M. LALANI, M. DE CHOUDHURY, A. SHARMA, AND N. KUMAR, *" like shock absorbers": Understanding the human infrastructures of technology-mediated mental health support*, in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–14.

[308] G. PENNYCOOK, T. D. CANNON, AND D. G. RAND, *Prior exposure increases perceived accuracy of fake news.*, Journal of experimental psychology: general, (2018).

[309] A. PERRIN, *Book reading 2016*, (2016).

[310] A. PERRIN AND M. ANDERSON, *Social media usage in the u.s. in 2019 | pew research center*.
https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/, 2019.
(Accessed on 09/08/2022).

[311] PESACHECK, *Pesacheck*.
https://pesacheck.org/, April 2021.
(Accessed on 04/15/2021).

[312] L. PETERSON, T. ANDERSON, D. CULLER, AND T. ROSCOE, *A blueprint for introducing disruptive technology into the internet*, SIGCOMM Computer Communication Review, 33 (2003), pp. 59–64.

[313] S. PHADKE AND T. MITRA, *Many faced hate: A cross platform study of content framing and information sharing by online hate groups*, in Proceedings of the 2020 CHI conference on human factors in computing systems, 2020, pp. 1–13.

[314] T. W. POST, *About the fact checker - the washington post*.
https://www.washingtonpost.com/politics/2019/01/07/about-fact-checker/, April 2021.
(Accessed on 04/15/2021).

[315] POYNTER, *The international fact-checking network*, accessed in March, 2021.

[316] PRERNA JUNEJA AND T. MITRA, *Algorithmic nudge: Using xai frameworks to design interventions*, CHI 2021 Workshop on Operationalizing Human-centered Perspectives in Explainable AI, (2021).

[317] PRERNA JUNEJA AND T. MITRA, *Human and technological infrastructures of fact-checking*, Proc. ACM Hum.-Comput. Interact., (2022).

[318] PRERNA JUNEJA, D. RAMA SUBRAMANIAN, AND T. MITRA, *Through the looking glass: Study of transparency in reddit's moderation practices*, Proc. ACM Hum.-Comput. Interact., 4 (2020).

[319] S. PULLAN AND M. DEY, *Vaccine hesitancy and anti-vaccination in the time of covid-19: A google trends analysis*, Vaccine, 39 (2021), pp. 1877–1881.

[320] K. PURCELL, *Findings: Search and email remain the top online activities | pew internet & american life project*, Pew Research Center's Internet & American Life Project, (2011).

[321] V. QAZVINIAN, E. ROSENGREN, D. R. RADEV, AND Q. MEI, *Rumor has it: Identifying misinformation in microblogs*, in Proceedings of the conference on empirical methods in natural language processing, Association for Computational Linguistics, 2011, pp. 1589–1599.

[322] T. QUINT, *Latest news, breaking news live, top news headlines, viral videos news updates - the quint*.
`https://www.thequint.com/`, April 2021.
(Accessed on 04/15/2021).

[323] L. RAINIE AND S. FOX, *The online health care revolution*, Pew Research Center, (2000).

[324] I. D. RAJI AND J. BUOLAMWINI, *Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products*, in Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 429–435.

[325] L. B. RASMUSSEN, *From human-centred to human-context centred approach: looking back over 'the hills', what has been gained and lost?*, Ai & Society, 21 (2007), pp. 471–495.

[326] T. N. REPUBLIC, *The new republic*.
`https://newrepublic.com/`, April 2021.
(Accessed on 04/15/2021).

[327] P. RESNICK, S. CARTON, S. PARK, Y. SHEN, AND N. ZEFFER, *Rumorlens: A system for analyzing the impact of rumors and corrections in social media*, in Proc. Computational Journalism Conference, vol. 5, 2014.

[328] M. REYNOLDS, *Amazon sells 'autism cure' books that suggest children drink toxic, bleach-like substances*, (2019).

[329] S. T. ROBERTS, *Behind the screen: The hidden digital labor of commercial content moderation*, PhD thesis, University of Illinois at Urbana-Champaign, 2014.

[330] R. E. ROBERTSON, S. JIANG, K. JOSEPH, L. FRIEDLAND, D. LAZER, AND C. WILSON, *Auditing partisan audience bias within google search*, Proceedings of ACM on Human Computer Interaction, 2 (2018), pp. 148:1–148:22.

[331] R. E. ROBERTSON, S. JIANG, K. JOSEPH, L. FRIEDLAND, D. LAZER, AND C. WILSON, *Auditing partisan audience bias within google search*, Proceedings of the ACM on Human-Computer Interaction, 2 (2018), pp. 1–22.

[332] R. E. ROBERTSON, D. LAZER, AND C. WILSON, *Auditing the personalization and composition of politically-related search engine results pages*, in Proceedings of the 2018 World Wide Web Conference, WWW '18, International World Wide Web Conferences Steering Committee, 2018, pp. 955–965.

[333] J. J. ROBINSON, J. MADDOCK, AND K. STARBIRD, *Examining the role of human and technical infrastructure during emergency response.*, in ISCRAM, 2015.

[334] S. RODDY, *Recent updates to amazon verified purchase reviews*, 2019.

[335] A. RODRIGUEZ, *Youtube's algorithms can drag you down a rabbit hole of conspiracies, researcher finds*, (2018).

[336] J. ROOZENBEEK AND S. VAN DER LINDEN, *Fake news game confers psychological resistance against online misinformation*, Palgrave Communications, 5 (2019), pp. 1–10.

[337] J. ROOZENBEEK AND S. VAN DER LINDEN, *Breaking harmony square: A game that "inoculates" against political misinformation*, The Harvard Kennedy School Misinformation Review, (2020).

[338] A. ROVETTA, *Infodemic emergency in italy: A longitudinal analysis of the web interest in sources of dis-misinformation, epidemiologically dangerous behaviors, and vaccine hesitancy during covid-19*, (2022).

[339] N. RUMMEL, H. SPADA, F. HERMANN, F. CASPAR, AND K. SCHORNSTEIN, *Promoting the coordination of computer-mediated interdisciplinary collaboration*, (2002).

[340] N. SAMBASIVAN AND T. SMYTH, *The human infrastructure of ictd*, in Proceedings of the 4th ACM/IEEE international conference on information and communication technologies and development, 2010, pp. 1–9.

[341] M. SAMORY AND T. MITRA, *Conspiracies online: User discussions in a conspiracy community following dramatic events*, in ICWSM, 2018.

[342] M. SAMORY AND T. MITRA, *'the government spies using our webcams': The language of conspiracy theories in online discussions*, Proceedings of the ACM on Human-Computer Interaction, 2 (2018), pp. 1–24.

[343] C. SANDVIG, K. HAMILTON, K. KARAHALIOS, AND C. LANGBORT, *An algorithm audit*, Data and Discrimination: Collected Essays. Washington, DC: New America Foundation, (2014), pp. 6–10.

[344] C. SANDVIG, K. HAMILTON, K. KARAHALIOS, AND C. LANGBORT, *Auditing algorithms: Research methods for detecting discrimination on internet platforms*, Data and discrimination: converting critical concerns into productive inquiry, 22 (2014).

[345] S. SAWYER AND A. TAPIA, *Always articulating: Theorizing on mobile and wireless technologies*, The Information Society, 22 (2006), pp. 311–323.

[346] N. SCHAROWSKI, *Transparency and Trust in AI*, PhD thesis, Institute of Psychology, 2020.

[347] A. B. SCHIFF, *090821_letter to amazon.pdf*. `https://schiff.house.gov/imo/media/doc/090821_Letter%20to%20Amazon.pdf`, month = Sep, year = 2021, note = (Accessed on 06/22/2023).

[348] P. SCHMIDT, F. BIESSMANN, AND T. TEUBNER, *Transparency and trust in artificial intelligence systems*, Journal of Decision Systems, 29 (2020), pp. 260–278.

[349] N. F. SCHNEIDEWIND, *The state of software maintenance*, IEEE Transactions on Software Engineering, (1987), pp. 303–310.

[350] G. SCHWITZER, *Pollution of health news*, 2017.

[351] S. SCUTTI, *Facebook to target vaccine misinformation with focus on pages, groups, ads*, (2019).

[352] S. SEARCHER, *Social searcher - free social media search engine*. https://www.social-searcher.com/, April 2021. (Accessed on 04/15/2021).

[353] A. SEITZ, *In election misinformation fight, '2020 changed everything' | ap news*. https://apnews.com/article/2022-midterm-elections-voting-rights-technology-business-social-media-f5ba340c7a98f6f058fb3afac74a26bb, 2022. (Accessed on 09/10/2022).

[354] J. C. M. SERRANO, O. PAPAKYRIAKOPOULOS, AND S. HEGELICH, *Nlp-based feature extraction for the detection of covid-19 misinformation videos on youtube*, in Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, 2020.

[355] V. SESSIONS AND M. VALTORTA, *The effects of data quality on machine learning algorithms.*, ICIQ, 6 (2006), pp. 485–498.

[356] C. SHAO, G. L. CIAMPAGLIA, A. FLAMMINI, AND F. MENCZER, *Hoaxy: A platform for tracking online misinformation*, in Proceedings of the 25th international conference companion on world wide web, 2016, pp. 745–750.

[357] G. SHEPARD, *The Real Watergate Scandal: Collusion, Conspiracy, and the Plot That Brought Nixon Down*, Simon and Schuster, 2015.

[358] J. SHIN AND T. VALENTE, *Algorithms and health misinformation: A case study of vaccine books on amazon*, Journal of Health Communication, (2020), pp. 1–8.

[359] P. SHIRALKAR, A. FLAMMINI, F. MENCZER, AND G. L. CIAMPAGLIA, *Finding streams in knowledge graphs to support fact checking*, in 2017 IEEE International Conference on Data Mining (ICDM), IEEE, 2017, pp. 859–864.

[360] J. SHUMWAY, *Oregon gop frontrunner for governor embraces claims of election fraud – oregon capital chronicle*. https://oregoncapitalchronicle.com/2022/02/01/oregon-gop-frontrunner-for-governor-embraces-claims-of-election-fraud/, 2022.

(Accessed on 09/08/2022).

[361] J. SIMKO, M. TOMLEIN, B. PECHER, R. MORO, I. SRBA, E. STEFANCOVA, A. HRCKOVA, M. KOMPAN, J. PODROUZEK, AND M. BIELIKOVA, *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, 2021, pp. 411–414.

[362] S. C. SIVEK AND S. BLOYD-PESHKIN, *Where do facts matter? the digital paradox in magazines' fact-checking processes*, Journalism Practice, 13 (2019), pp. 998–1002.

[363] A. SMITH, V. KUMAR, J. BOYD-GRABER, K. SEPPI, AND L. FINDLATER, *Closing the loop: User-centered design and evaluation of a human-in-the-loop topic modeling system*, in 23rd International Conference on Intelligent User Interfaces, 2018, pp. 293–304.

[364] P. L. SOO RIN KIM, LAURA ROMERO AND K. HOLLAND, *With 10 weeks until midterms, election deniers are hampering some election preparations - abc news*.
https://abcnews.go.com/US/10-weeks-midterms-election-deniers-hampering-election-preparations/story?id=89007798, 2022.
(Accessed on 09/07/2022).

[365] A. SPAA, A. DURRANT, C. ELSDEN, AND J. VINES, *Understanding the boundaries between policymaking and hci*, in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, New York, NY, USA, 2019, Association for Computing Machinery, p. 1–15.

[366] SPOONBILL, *Spoonbill*.
https://spoonbill.io/, April 2021.
(Accessed on 04/15/2021).

[367] S. L. STAR AND K. RUHLEDER, *Steps toward an ecology of infrastructure: Design and access for large information spaces*, Information systems research, 7 (1996), pp. 111–134.

[368] S. L. STAR AND A. STRAUSS, *Layers of silence, arenas of voice: The ecology of visible and invisible work*, Computer supported cooperative work (CSCW), 8 (1999), pp. 9–30.

[369] M. STEIGER, T. J. BHARUCHA, S. VENKATAGIRI, M. J. RIEDL, AND M. LEASE, *The psychological well-being of content moderators*, (2021).

[370] M. STENCEL, *Number of fact-checking outlets surges to 188 in more than 60 countries*, 2019.

[371] H. STEPNICK, *How will social media platforms respond to election misinformation? it isn't clear - poynter*.
https://www.poynter.org/fact-checking/2022/how-will-social-media-platforms-respond-to-election-misinformation-it-isnt-clear/, 2022.
(Accessed on 09/08/2022).

[372] A. STISEN, N. VERDEZOTO, H. BLUNCK, M. B. KJÆRGAARD, AND K. GRØNBÆK, *Accounting for the invisible work of hospital orderlies: Designing for local and global coordination*, in Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, 2016, pp. 980–992.

[373] N. J. STROUD, *Polarization and partisan selective exposure*, Journal of communication, 60 (2010), pp. 556–576.

[374] C. R. SUNSTEIN, *Conspiracy theories and other dangerous ideas*, Simon and Schuster, 2014.

[375] C. TANG, Y. CHEN, B. C. SEMAAN, AND J. A. ROBERSON, *Restructuring human infrastructure: The impact of ehr deployment in a volunteer-dependent clinic*, in Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, 2015, pp. 649–661.

[376] Y. R. TAUSCZIK AND J. W. PENNEBAKER, *The psychological meaning of words: Liwc and computerized text analysis methods*, Journal of language and social psychology, 29 (2010), pp. 24–54.

[377] N. TAYLOR, K. CHEVERST, P. WRIGHT, AND P. OLIVIER, *Leaving the wild: lessons from community technology handovers*, in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2013, pp. 1549–1558.

[378] A. O. A. M. TEAM, *45% of american adults doubt vaccine safety, according to survey*, (2019).

[379] T. Y. TEAM, *Continuing our work to improve recommendations on youtube*, (2019).

[380] D. TEYSSOU, J.-M. LEUNG, E. APOSTOLIDIS, K. APOSTOLIDIS, S. PAPADOPOULOS, M. ZAMPOGLOU, O. PAPADOPOULOU, AND V. MEZARIS, *The invid plug-in: web video verification on the browser*, in Proceedings of the first international workshop on multimedia verification, 2017, pp. 23–30.

[381] THE WASHINGTON POST, *Fact checker - the washington post*. https://www.washingtonpost.com/news/fact-checker/, April 2021. (Accessed on 04/15/2021).

[382] THE YOUTUBE TEAM, *Managing harmful conspiracy theories on youtube*. https://blog.youtube/news-and-events/harmful-conspiracy-theories-youtube/, 2020. (Accessed on 09/08/2022).

[383] L. THOMAS, *Pins and needles: Pinterest tackles spread of vaccine misinformation*, (2019).

[384] A. THOMPSON, *Trump deploys youtube as his secret weapon in 2020 - politico*. https://www.politico.com/news/2020/09/06/trumpyoutube-election-comeback-408576, 2020. (Accessed on 09/08/2022).

[385] J. THORNE, A. VLACHOS, C. CHRISTODOULOPOULOS, AND A. MITTAL, *Fever: a large-scale dataset for fact extraction and verification*, arXiv preprint arXiv:1803.05355, (2018).

[386] L. A. TIMES, *Man inspired by false 'pizzagate' rumor on internet pleads guilty to shooting at d.c. restaurant*, (2017).

[387] T. N. Y. TIMES, *The new york times/cbs news poll*, (2004).

[388] D. TINGLEY AND G. WAGNER, *Solar geoengineering and the chemtrails conspiracy on social media*, Palgrave Communications, 3 (2017), p. 12.

[389] P. TOLMIE, R. PROCTER, D. W. RANDALL, M. ROUNCEFIELD, C. BURGER, G. WONG SAK HOI, A. ZUBIAGA, AND M. LIAKATA, *Supporting the use of user generated content in journalistic practice*, in Proceedings of the 2017 chi conference on human factors in computing systems, 2017, pp. 3632–3644.

[390] M. TOMLEIN, B. PECHER, J. SIMKO, I. SRBA, R. MORO, E. STEFANCOVA, M. KOMPAN, A. HRCKOVA, J. PODROUZEK, AND M. BIELIKOVA, *An audit of misinformation filter bubbles on youtube: Bubble bursting and recent behavior changes*, in Fifteenth ACM Conference on Recommender Systems, 2021, pp. 1–11.

[391] J. M. P. TORNERO, S. S. TAYIE, S. TEJEDOR, AND C. PULIDO, *How to confront fake news through news literacy? state of the art.*, Doxa Comunicación, (2018).

[392] TRELLO, *Trello*.
`https://trello.com/en-US`, April 2021.
(Accessed on 04/15/2021).

[393] D. TRIELLI AND N. DIAKOPOULOS, *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, CHI '19, ACM, 2019, pp. 453:1–453:15.

[394] D. TRIELLI AND N. DIAKOPOULOS, *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, 2019, pp. 1–15.

[395] D. TUFFLEY, *Mind the gap*, International Journal of Sociotechnology and Knowledge Development (IJSKD), 1 (2009), pp. 58–69.

[396] K. T. UNRUH AND W. PRATT, *The invisible work of being a patient and implications for health care:"[the doctor is] my business partner in the most important business in my life, staying alive."*, in Ethnographic Praxis in Industry Conference Proceedings, vol. 2008, Wiley Online Library, 2008, pp. 40–50.

[397] E. VAN COUVERING, *Is relevance relevant? market, science, and war: Discourses of search engine quality*, Journal of Computer-Mediated Communication, 12 (2007), pp. 866–887.

[398] S. VAN DER LINDEN, *Misinformation: susceptibility, spread, and interventions to immunize the public*, Nature Medicine, 28 (2022), pp. 460–467.

[399] T. G. VAN DER MEER AND Y. JIN, *Seeking formula for misinformation treatment in public health crises: The effects of corrective information type and source*, Health Communication, 35 (2020), pp. 560–575.

[400] M. VAN DER MEULEN AND W. G. REIJNIERSE, *Factcorp: A corpus of dutch fact-checks and its multiple usages*, in Proceedings of The 12th Language Resources and Evaluation Conference, 2020, pp. 1286–1292.

[401] N. VERDEZOTO, N. BAGALKOT, S. Z. AKBAR, S. SHARMA, N. MACKINTOSH, D. HARRINGTON, AND P. GRIFFITHS, *The invisible work of maintenance in community health: Challenges and opportunities for digital health to support frontline health workers in karnataka, south india*, Proceedings of the ACM on Human-Computer Interaction, 5 (2021), pp. 1–31.

[402] G. VERMA, A. BHARDWAJ, T. ALEDAVOOD, M. DE CHOUDHURY, AND S. KUMAR, *Examining the impact of sharing covid-19 misinformation online on mental health*, Scientific Reports, 12 (2022), pp. 1–9.

[403] R. VERN AND S. K. DUBEY, *Evaluating the maintainability of a software system by using fuzzy logic approach*, Int. J. Information Technology and Computer Science, 7 (2014), pp. 67–72.

[404] Y. VIEWERS, *Learn about watch history on youtube - youtube*. `https://www.youtube.com/watch?v=YbWZcgOYHAc&ab_channel=YouTubeViewers`, 2022.
(Accessed on 02/07/2022).

[405] N. VINCENT, B. HECHT, AND S. SEN, *"data strikes": Evaluating the effectiveness of a new form of collective action against technology companies*, in The World Wide Web Conference, 2019, pp. 1931–1943.

[406] N. VINCENT, I. JOHNSON, P. SHEEHAN, AND B. HECHT, *Measuring the importance of user-generated content to search engines*, Proceedings of the International AAAI Conference on Web and Social Media, 13 (2019), pp. 505–516.

[407] N. VINCENT, I. JOHNSON, P. SHEEHAN, AND B. HECHT, *Measuring the importance of user-generated content to search engines*, in Proceedings of the International AAAI Conference on Web and Social Media, vol. 13, 2019, pp. 505–516.

[408] J. VITAK, P. ZUBE, A. SMOCK, C. T. CARR, N. ELLISON, AND C. LAMPE, *It's complicated: Facebook users' political participation in the 2008 election*, CyberPsychology, behavior, and social networking, 14 (2011), pp. 107–114.

[409] D. WAKABAYASHI, *Election misinformation continues staying up on youtube. - the new york times*.
https://www.nytimes.com/2020/11/10/technology/election-misinformation-continues-staying-up-on-youtube.html, Accessed on 09/08/2022.

[410] W. Y. WANG, *" liar, liar pants on fire": A new benchmark dataset for fake news detection*, arXiv preprint arXiv:1705.00648, (2017).

[411] X. WANG, N. GOLBANDI, M. BENDERSKY, D. METZLER, AND M. NAJORK, *Position bias estimation for unbiased learning to rank in personal search*, in Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, ACM, 2018, pp. 610–618.

[412] B. WASSON, *Identifying coordination agents for collaborative telelearning*, International Journal of Artificial Intelligence in Education (IJAIED), 9 (1998), pp. 275–299.

[413] W. WEBBER, A. MOFFAT, AND J. ZOBEL, *A similarity measure for indefinite rankings*, ACM Transactions on Information Systems (TOIS), 28 (2010), pp. 1–38.

[414] M. WEBSTER, *Stringer definition & meaning - merriam-webster*.
https://www.merriam-webster.com/dictionary/stringer, 2022.
(Accessed on 04/17/2022).

[415] C. G. WEISSMAN, *Despite recent crackdown, youtube still promotes plenty of conspiracies*, (2019).

[416] C. G. WEISSMAN, *Despite recent crackdown, youtube still promotes plenty of conspiracies*, (2019).

[417] J. WHITTAKER, S. LOONEY, A. REED, AND F. VOTTA, *Recommender systems and the amplification of extremist content*, Internet Policy Review, 10 (2021), pp. 1–29.

[418] WHOIS.NET, *Whois lookup & ip | whois.net*.
https://www.whois.net/, April 2021.
(Accessed on 04/15/2021).

[419] WIKIPEDIA, *Conspiracy theory*, (2002).

[420] WIKIPEDIA, *Malaysia airlines flight 370*, (2019).

[421] WIKIPEDIA, *Project mkultra*, (2019).

[422] WIKIPEDIA CONTRIBUTORS, *2009 swine flu pandemic*, (2020).

[423] C. WILSON, *The promise and peril of algorithm audits for increasing transparency and accountability of donated datasets*, (2019).

[424] S. WINEBURG AND S. MCGREW, *Lateral reading: Reading less and learning more when evaluating digital information*, (2017).

[425] M. WOOD, *Has the internet been good for conspiracy theorising*, PsyPAG Quarterly, 88 (2013), pp. 31–34.

[426] WORLD HEALTH ORGANIZATION, *Six common misconceptions about immunization*, (2019).

[427] Z. WU, Y. LIU, Q. ZHANG, K. WU, M. ZHANG, AND S. MA, *The influence of image search intents on user behavior and satisfaction*, in Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, ACM, 2019, pp. 645–653.

[428] W. XIAO, H. ZHAO, H. PAN, Y. SONG, V. W. ZHENG, AND Q. YANG, *Beyond personalization: Social content recommendation for creator equality and consumer satisfaction*, in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 235–245.

[429] X. XIE, Y. LIU, M. DE RIJKE, J. HE, M. ZHANG, AND S. MA, *Why people search for images using web search engines*, in Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, ACM, 2018, pp. 655–663.

[430] D. XIN, L. MA, J. LIU, S. MACKE, S. SONG, AND A. PARAMESWARAN, *Accelerating human-in-the-loop machine learning: Challenges and opportunities*, in Proceedings of the second workshop on data management for end-to-end machine learning, 2018, pp. 1–4.

[431] YOUGOV, *Most flat earthers consider themselves very religious*, (2018).

[432] D. G. YOUNG, K. H. JAMIESON, S. POULSEN, AND A. GOLDRING, *Fact-checking effectiveness as a function of format and tone: Evaluating factcheck. org and flackcheck. org*, Journalism & Mass Communication Quarterly, 95 (2018), pp. 49–75.

[433] YOUTUBE, *Supporting the 2020 u.s. election*, (2020).

[434] ——, *Youtube community guidelines*, (2020).

[435] YOUTUBE, *View or delete search history - computer - youtube help*. https://support.google.com/youtube/answer/57711?co=GENIE.Platform%3DDesktop&hl=en, 2022. (Accessed on 02/07/2022).

[436] YOUTUBE, *Elections misinformation policies - youtube help*. https://support.google.com/youtube/answer/10835034?hl=en, Accessed on 09/08/2022.

[437] YOUTUBE, *Browse youtube while incognito on mobile devices - youtube help*, (Accessed on 09/14/2022).

[438] Q. ZHANG, A. LIPANI, S. LIANG, AND E. YILMAZ, *Reply-aided detection of misinformation via bayesian deep learning*, in The world wide web conference, 2019, pp. 2333–2343.

[439] M. ZONIS AND C. M. JOSEPH, *Conspiracy thinking in the middle east*, Political Psychology, (1994), pp. 443–459.