CS333 Final Project: The Political Implications of the YouTube Recommendation System

Nathan Kim (njk24), Jeffrey Luo (jl834), Rebecca Shu (rjs73), Elizabeth Zhang (eyz3)

Introduction and Context

From its creation in 2005, the video streaming and sharing application, YouTube, has seen consistent growth and user engagement, earning the accolade of being the second most visited website worldwide, as of 2021. With more than one billion monthly users, YouTube has had a substantial social impact in influencing popular culture, contributing to internet trends, and creating celebrities and influencers. However, YouTube has also been criticized for facilitating the spread of misinformation, violating its users' privacy, enabling censorship, and endangering child safety and wellbeing. Specifically, there has been much controversy surrounding YouTube's recommendation system, in which curated videos are recommended to each individual user based on the user's history of watched content, and the videos they have had a history of liking and interacting with. Each personalized content feed or advertisement placement is carefully curated to the user through this recommendation algorithm, and due to each user's unique viewing habits, certain content may be filtered out in place of content that is more in-line with what the user has had history of liking before.

This can be dangerous in many aspects, including when looking at these systems from a political lens. Due to the lack of regulation as well as the recommendation system, YouTube has been criticized for spreading misinformation regarding political news channels, government conspiracy theories, radicalized ideas, and more recently, misinformation about COVID-19 and the vaccine. Many of these ideas have political links, and have even led to the outbreak of violence, racism, and other hate crimes. Therefore, we have chosen to explore the impact of YouTube's recommendation system, and its political implications on aspects of user's everyday lives.

Algorithms

There are two main types of algorithms for recommendation systems in general: collaborative filtering and content-based filtering.

Collaborative methods for recommender systems are methods that are based solely on the past interactions recorded between users and items in order to produce new recommendations. These interactions are stored in the so-called "user-item interactions matrix". Then, upon being stored in these matrices, these past user-item interactions are sufficient to detect similar users and/or similar items and make predictions based on these estimated proximities. Collaborative filtering algorithms can be further divided into two sub-categories that are generally called memory based and model based approaches. Memory based approaches directly work with values of recorded interactions, assuming no model, and are essentially based on nearest neighbors search (for example, find the closest users from a user of interest and suggest the most popular items among these neighbors). Model based approaches take the recorded user-item interactions and try to discover an underlying "reasoning" model that explains the user-item interactions. After discovering such a model, the algorithm then takes the discovered model to make predictions.

The second type of recommendation system is content-based filtering. Content-based filtering differs from collaborative filtering in that instead of only relying on user-item interactions, content-based approaches use additional information about users and items. Characteristics like age, sex, job, or other personal information for users might be accounted for in this kind of algorithm. The characteristics of the items would also be additional information. The basic process performed by a content-based recommender consists of matching up the characteristics of a user profile in which preferences and interests are stored, with the characteristics of an item, in order to recommend to the user new interesting items.

Focusing on YouTube, its recommendation system has evolved greatly since its introduction in 2005. In its early years, the YouTube algorithm recommended the videos that attracted the most views or clicks. Unfortunately, this led to a proliferation of misleading titles and thumbnails,

otherwise known as clickbait. User experience declined as videos left people feeling tricked, unsatisfied, or simply annoyed. In 2012, YouTube adjusted its recommendation system to support time spent watching each video, as well as time spent on the platform overall. This led some creators to try to make their videos shorter in order to make it more likely viewers would watch to completion, whereas others made their videos longer in order to increase watch time overall. YouTube didn't endorse either of these tactics, and maintained the stance: make high quality videos, and the algorithm will reward you. YouTube soon realized that long times spent on the platform is not necessarily equivalent to quality time spent.

In 2015, YouTube began measuring viewer satisfaction directly with user surveys as well as prioritizing direct response metrics like Shares, Likes and Dislikes. In 2016, YouTube pushed out the preliminary stages of the algorithm it now uses. YouTube recommendations are driven by Google Brain, which was recently opensourced as TensorFlow (Covington, 2016). The system consists of two neural networks (Covington, 2016). The first one, candidate generation, takes as input user's watch history and using collaborative filtering selects videos in the range of hundreds. The second neural network is used for ranking the few hundreds of videos in order. This is much simpler as a problem than candidate generation as the number of videos is smaller and more information is available for each video and its relationship with the user. This system uses logistic regression to score each video and then A/B testing is continuously used for further improvement.

Research questions

Recommendation systems are used in almost all social media outlets, influencing people's opinions and beliefs on a daily basis, including YouTube. Through their platform, they control the information we engage with by filtering and pushing certain videos with viewpoints geared toward specific audiences, shaping each individual's access to information and molding people's perspectives on topics like politics. Our goal is to research the implications of YouTube's algorithm on people's political beliefs and how the spread of misinformation can contribute to radicalization and political polarization. YouTube's algorithm can potentially create personalized feedback loops that continue to push the same political dialogues, creating a bubble that leads to biases and extreme viewpoints. Oftentimes, videos related to political conspiracy theories can lead to echo chambers that then creates never-ending rabbit holes that one can spend hundreds of hours in. Therefore, the questions we hope to answer are: *To what extent does the YouTube recommendation algorithm impact political polarization and biases? What are the implications of the role of YouTube's spread of misinformation and extremist content on users' behaviors and political stances?*

Our research questions are feasible and substantial because they hone in on a specific platform, YouTube, and the implications of YouTube's algorithm on users' political beliefs that we can further dive into with the resources addressed in the methods section, which we have broken down into sections for each group member to analyze deeper. These questions are substantial because they involve more than a surface level analysis; they require us to have a deep understanding of the algorithm, as well as the specific implications of the algorithm on a user's behavior and beliefs. Furthermore, because YouTube is one of the dominating video streaming services, their content is viewed by hundreds of millions of people daily. The scope of their platform gives them enormous power, creating effects that can influence the political climate of a country, which we studied from past elections and political events. There is an abundance of research done on YouTube with differing opinions and articles on this topic, which we put in substantial effort to sift through to find credible and reliable sources.

Results & Methods

To what extent does the YouTube recommendation algorithm impact political polarization and biases?

YouTube as a platform inherently relies on a viewer's retention time on the website in order to gain revenue and have a stronger presence on the internet. As a result of this, YouTube's algorithm takes into account watch history, search history, and retention time on videos in order to recommend videos that would help keep users on their site. This user-catered experience has led many to question

the recommendation algorithm's role in creating politically polarized communities on the platform. Our research has found that because users are able to selectively seek information that aligns with their political beliefs, the personalization of the videos they engage in tend to create filter bubbles and echo chambers that create homogeneous environments, perpetuating the same ideas which further segregate ideologies through the content that aligns with the user's pre-existing beliefs. Although there is no direct evidence that the YouTube algorithm single-handedly increases political polarization, there is sufficient research that shows how other phenomena like feedback loops, echo chambers, and filter bubbles as well as misinformation can reinforce a user's existing beliefs which then polarizes the community's perspectives.

Political Polarization: Echo Chambers and Filter Bubbles

Polarization in many democratic societies is linked to the rise in new forms of communication and social media. The personalization of media allows users to select their own content which leads to filtering bubbles that segregate ideologies through the recommending of content that align with the user's pre-existing beliefs. Recommender algorithms usually base selections on a user's personal search history (content-based filtering) or the interactions between users and items (collaborative filtering). The user's actions reveal their interests and preferences and feed the algorithm information to personalize their feed. This algorithmic personalization can shape information consumption and lead to filter bubbles that create environments that only consist of similar opinions and political preferences which further segregates users by their ideological and political views.

In a study published in the Journal of Broadcasting & Electronic Media, researchers manipulate user search/watch history to see if algorithmically recommended content reinforces and polarizes political opinions. A week after the 2016 U.S. presidential election, 108 undergraduate university students were shown 26 terms/statements about Hillary Clinton and Donald Trump and were told to rank the top 10 according to personal interest. They were also asked to rank the top 10 statements that they would see posted/recommended on their friends' feeds. A list of search terms was then created and fed to a new YouTube account for each participant, and the videos from the search items as well as recommended videos were played to bias the search/watch history for each YouTube account. The first 5 recommended videos based on the user's own preferences or the expected preferences of their social circle were then shown to each participant with the control being an unbiased account searching "the 2016 Presidential Election". Participants completed a questionnaire assessing their emotional attitudes toward each party candidate before and after watching the videos, and results showed that users' emotion-ideology alignment was heightened by the political videos recommended to them through the algorithm based on their own search history (Cho et al., 2020). Thus, when algorithmic recommendation systems are personalized by users' data, they can strengthen political beliefs, creating an echo chamber that reinforces users' pre-existing beliefs and encourages polarized opinions. However, when looking at the results from the search terms based on what they believe their social circle would engage in, the reinforcement patterns were less significant. This shows that selective exposure is key in creating echo chambers that reinforce these behaviors, and if the algorithm includes a broader range of search terms, there is potential for users to explore a wider variety of viewpoints, lessening the political selectivity encouraged by the algorithms.

When looking at political communication in our society, we see that it consists of politicians, information mediators, and citizens, where politicians provide the source of the information which the media then acts as a mediator for the citizens to receive the information. The media can have different effects on people's behaviors and judgment, and there is evidence suggesting that social media can potentially cause an increase in political polarization. Oftentimes, media consumption can be transactional in the sense that users choose what information they engage with, leading to an increase in personalization of media. This algorithmic personalization "lies at the core of the 'demassification of mass communication' because it further allows media users to select their own media content", and this personalization is then used to increase engagement to keep users hooked (García-Marín, J., 2021). However, that is not always the case across different studies because of the lack of mediators in certain social media platforms. In a study published in the book series *Studies in Digital Politics and Governance*, they specifically focus on the polarization of the 2017 independence crisis in Catalonia based on the analysis of the polarization in traditional media compared to the videos on YouTube and

whether or not the absence of a moderator between the target audience and information source contributes to polarization. The study analyzed 10,000 news stories, 193 videos, and 80,000+ comments through sentiment analysis based on the Gentzkow Polarization by calculating the distance between the negative and positive sentiment between the two sides of the political spectrum. From the results, the researchers found that the information from YouTube showed greater extremes of polarization than traditional media with the dispersion of the polarization variable at a greater variability and significant outliers compared to those of traditional media. However, amongst the traditional media, there were differing degrees of polarization which was divided into two groups of either mostly print or digital media outlets, highlighting the effects of lower cost media that can rely on more segmented audiences. Researchers concluded that social networks cannot be entirely blamed for the increase in polarization, but instead they require polarizing agents such as social media platforms to create environments that function as filter bubbles and echo chambers in order to affect and increase the political polarization on a platform. However, traditional media does not have a filter or mediator between the source of the information and the citizens receiving the information which is why YouTube showed more polarization. Without a mediator that strives to provide citizens information they are looking for, social media may not always have a polarizing effect which can explain the differences found in previous studies.

Additionally, news is consumed through non-algorithmic sources, user-driven algorithmically generated sources, and socially driven algorithmically generated sources. YouTube is defined as a user-driven news source because it takes into account a user's demographics, location, and watch/search history to develop a series of recommended videos. This will most likely encourage homogenization through the narrow selection of perspectives in videos recommended that reinforce the same beliefs. However despite previous research that suggests these claims, the findings from a study published in the Computers in Human Behavior Journal only showed an increase in political participation from user-driven and socially driven algorithmically generated sources. Research shows that individuals will pay more attention to negative news, focusing on the negative aspects of opposing candidates to reaffirm their support for their preferred candidate. People also tend to pay more attention to news more relevant or aligned with their beliefs. Because user-driven algorithmic sources are more likely to reinforce negative emotions like fear and anxiety which studies have shown motivate participation, users who consume content from user-driven sources are more likely to participate. Because algorithms want to maximize user attention, user-driven algorithms have no incentive to provide counter-attitudinal information unless the user specifically clicks on it. We would expect that user-driven algorithmic sources would lead to more polarization, but results from a study done on the analysis of a sample of young adults showed that algorithmic news sources only had a significant effect on traditional and online political participation. These findings suggest that social media platforms have the potential to "facilitate these phenomena such as dislike or negative partisanship, but may not, on their own, be the source of polarizing beliefs" (Feezell et al., 2021). This study shows that the algorithm may not solely be responsible for increasing political polarization, but instead the occurrences of filter bubbles and echo chambers may have a larger effect on polarization.

The emphasis on the importance of echo chambers and filter bubbles is once again highlighted in research studies done on the polarization of users, leading us to believe that the algorithm itself may not be the only factor. Because users will selectively choose to expose themselves to information that aligns with their beliefs, this confirmation bias then forms polarized groups, creating echo chambers. However, studies show that the algorithm is not single handedly responsible in creating the increase in these polarization viewpoints. When comparing how user behaviors are affected when exposed to the same content on different platforms like YouTube and Facebook, one research study is able to look at the polarization induced not only by the platform's algorithm but also by the content promoted. When researchers looked into the polarization dynamics and how users interacted with different types of content, they noticed that even if users start off with commenting on both sides, they eventually focus on one main narrative. This experiment uses science and conspiracy videos because they inherently have conflicting narratives and are on the two ends of the spectrum. By observing how users distribute their comments on science vs conspiracy news posts on both platforms, they found that users usually concentrated their actions on one of the extreme narratives. This aggregation of users around a certain narrative then leads to homogenous polarized groups forming. Through these findings, researchers have concluded that polarization may depend on

the content more than the algorithm promoting the content. Even looking at the polarized users on each platform, the users interact with the extreme content they support similarly. On both Facebook and YouTube platforms, researchers have found that polarized and homogeneous communities emerge and behave similarly despite different algorithmic platforms. Researchers have always focused on the algorithm itself for causing polarization effects through social media. However, when looking at similar content across different platforms, results show that "conflicting narratives lead to the aggregation of users in homogenous echo chambers, irrespective of the online social network and the algorithm" (Hussein et al., 2020). Regardless of the algorithm, homogeneous echo chambers emerge, creating similar polarizing behaviors which leads us to believe that we should not only focus on the harmful effects of the algorithm, but also the content YouTube itself is allowing.

The Spread of Conspiracy Theories and Rabbit Holes

A large topic of contention when looking at YouTube's role in the political sphere is the existence of conspiracy theory videos and the potential for rabbit holes. Especially when considering the rising popularity of conspiracy theories related to more contemporary topics related to TV and cinema, does YouTube have a role in promoting politically-driven conspiracy theory videos and rabbit holes?

When considering that the goal of most recommender algorithms is to maximize an individual user's time on the platform, it is easy to become skeptical of YouTube's role in political radicalization. One term to be familiar with when analyzing recommender systems is "technological seduction". Coined in the paper "Technologically scaffolded atypical cognition: The case of YouTube's recommender system", it refers to the use of predictive analytics to make a user-catered experience. For YouTube, this is done by using factors such as geolocation, search history, watchtime retention, etc. in order to interfere with the reasoning process and thus 'seduce' the user into spending more time on the platform (Alfano et al., 2020). This study goes on to use a YouTube data crawler to create seed terms that are then used to test whether YouTube's recommender system is liable to promote and perhaps intensify conspiracy theories, as these viewpoints have the most potential to cause social harm. The paper found that accounts with activity revolving around the topic of "gurus" on YouTube were the most susceptible to being recommended conspiracy theories, but that topics that are even more unrelated to conspiracy theories like "fitness" and "natural foods" were still somehow susceptible to being recommended conspiracy theory videos. These findings do point to some consistent level of promotion for conspiracy theory videos, no matter how unrelated a user's interests and watch history are from the topic.

This notion that YouTube does tend to lean towards promoting conspiracy theory videos and rabbit holes is supported by the study "Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems". This study used YouTube's co-visitation counts and prior viewing history (factors in technological seduction) that are used to provide recommendations to users. The authors primarily focused on developing categories to test and confirm what kind of videos were recommended to the average viewer on YouTube. They generally found that it is very possible for "a user to be immersed in this content following a short series of clicks." (O'Callaghan et al., 2014). The article as a whole also highlights that more attention needs to be drawn to how these recommender systems on all major platforms like Twitter, YouTube, Amazon, etc. can have unexplored ways in which political thought and action can be formed and promoted.

However, there are some cases where it seems as though YouTube tries to minimize the presence of conspiracy theory videos on their site, particularly when these theories or ideas can potentially be deemed 'dangerous'. The paper '"Down the rabbit hole" of vaccine misinformation on YouTube: Network exposure study' used networks of videos based on YouTube recommendations regarding pro-vaccine and anti-vaccine keywords, known as goal-oriented browsing, as well as conspiracy and anti-vaccine expert seed videos, known as direct navigation (Tang et al., 2021) . Through these methods, the authors were able to find that viewers are much more likely to be recommended anti-vaccine content using direct navigation than using goal-oriented browsing. Thus, the authors were able to conclude that YouTube artificially pushes the popularity and exposure of pro-vaccine videos in order to mitigate the dangers of misinformative anti-vaccine videos. However, the inclination of the algorithm itself is to further push anti-vaccine content to viewers when brought

onto an anti-vaccine video from another website, thus creating deeper and deeper rabbit holes and the likelihood for vaccine related conspiracy theories to enter the "echo chamber" in terms of viewership. This shows that YouTube does indeed play a role in suppressing potentially dangerous conspiracies despite the algorithm's inclination to promote such content.

All of these findings highlight the psychological factors that make YouTube's algorithm effective and even possible. According to "Differential Susceptibility to Misleading Flat Earth Arguments On YouTube", there are psychological implications of even being on a platform like YouTube, in which "individuals are biased in how they search, attend to, evaluate, and incorporate new information into their existing knowledge, often privileging and prioritizing information that aligns with their pre-existing views and values" (Landrum et al., 2021) . This selective exposure along with the user-oriented recommendation system strengthens the viewer's inclination to watch content that they are already interested in. This study aims to take into account selective exposure and a user's susceptibility to certain types of media based on certain conditions like media use, individual susceptibility variables, response states, and media effects in order to identify which dispositions and factors contribute to susceptibility towards flat Earth beliefs. The authors conclude that although flat Earth content on YouTube does not convince the majority of viewers, there are those that are more open to believing in the flat Earth conspiracy, but those that are already inclined towards these views or have low scientific background information can be easily swayed towards being recommended more videos that echo the same beliefs. This brings to light that the recommendations made on the YouTube platform are effective not on the algorithm alone, but rather the inclination of viewers to selective exposure.

What are the implications of the role of YouTube's spread of misinformation and extremist content on users' behaviors and political stances?

With free speech comes unverified information, misinformation, and conspiracy theories that are often uncredible. Due to the advertising nature of YouTube's recommendation algorithm, from a monetary standpoint, YouTube's recommendation and search algorithms do not necessarily need to consider a piece of content's accuracy in its recommendation decisions, as long as the targeted audience provides feedback that the content is welcome. In addition, due to the sheer amount of content, millions of hours of videos uploaded, and billions of viewers that YouTube reaches, there exists a need to rely on a computerized recommendation algorithm in order to scale to and control each user's personalized video feeds. However, in the failure to control the amount of misinformation, graphic, and sometimes extremist content present, YouTube becomes indirectly responsible for this spreading of inappropriate content. Our research has shown that only in situations in which there has been public pressure for YouTube to reform its controlling of content, such as in the case of the spreading of fake COVID-19 vaccine news as well as the call to reform content in order to combat extremist content online, has YouTube made active efforts and has seen progress made towards the regulation of potentially dangerous content, in those areas regarding the COVID-19 vaccination and terrorist content specifically. When YouTube has put in substantial effort and work towards these efforts to minimize extremist and misleading content, research has shown that the amount of misinformative content has decreased substantially, and viewer's public sentiment has been steered more towards scientifically correct content. However, there is still copious amounts of misinformation and dangerous content that exists on the platform today. Therefore, YouTube's recommendation system needs to be reformed to become proactive instead of reactive, and an active effort needs to be made to control the spread of other misinformation and extremist content, beyond what is just publicly called for.

Radicalization

When exploring the implications of the role of YouTube's spread of misinformation and extremist content on users' behaviors and political stances, it is important to first understand whether YouTube's recommendation algorithm contributes, or does not contribute, to radicalization or other extremist views. These issues have had wide-spread implications on the safety and stance of people's

views on government issues, and the dangers of radicalization have historically led to acts of violence, terrorism, and other extremist acts and beliefs.

The role that YouTube and its behind-the-scenes recommendation algorithm plays in encouraging online radicalization has been an idea that has long been suggested by journalists, media, and other members in academia. YouTube channels that discuss social, political, and cultural subjects have often flourished on the platform, and frequently, the posted videos focus on controversial topics such as race, gender, and religion. The people who post such videos span a wide range of political orientations, from podcast hosts like Joe Rogan to advocates of white supremacy, like Richard Spencer (Ribeiro, 2020). Because these people post on the same platform, interactions between these vastly different people on the website is inevitable. For example, Joe Rogan interviewed YouTuber Carl Benjamin, who debated with white supremacist Richard Specner. This kind of connectivity may create “radicalization pathways" for audience members and content creators. Due to YouTube’s extreme popularity, if the streaming website is actually radicalizing individuals, this could push extremist ideologies like white supremacy into the mainstream.

However, in their paper “Algorithmic Extremism: Examining YouTube’s Rabbit Hole of Radicalization”, researchers found that YouTube’s recommendation algorithm fails to promote inflammatory or radicalized content, as previously claimed by several outlets. In the study, researchers categorized nearly 800 political channels and “were able to differentiate between political schemas in order to analyze the algorithm traffic flows out and between each group” (Ledwich & Zaitsev, 2020). Researchers were then able to conduct a detailed analysis of the recommendations that were received by each channel type. Upon gathering this information, it was found that contrary to popular belief, the data suggested that YouTube’s recommendation algorithm actually actively discourages viewers from visiting radicalizing or extremist content. In fact, the algorithm has been shown to instead favor more mainstream media and cable news content, rather than independent YouTube channels that were left-leaning or politically neutral channels. Furthermore, though researchers concluded that there was more right-wing content present on YouTube than ever before, “these more numerous channels gained only a fraction of the views of mainstream media and centrist channels” (Ledwich & Zaitsev, 2020). Specifically, videos categorized as being center or left leaning in content had 22 million daily users, while videos categorized as being right-wing had only 5.6 million daily users. Data from this study showed that even if a user was watching very extreme content, their recommendations would be populated with a mixture of extreme and more mainstream content. YouTube is, therefore, more likely to steer people away from extremist content rather than vice versa.

Similarly, in “Examining the consumption of radical content on YouTube”, researchers analyzed the news consumption and browsing behaviors of over 300,000 YouTube users, who had watched almost half a million of videos related to politics over a 4 year period. Researchers were able to study 309,813 users who had at least one recorded YouTube pageview. They then found a total of 21,385,962 watched-video pageviews, and were able to quantify the user’s attention by the duration of in-focus visit to each video in total minutes. As each YouTube video has a unique identifier embedded in its URL, researchers were able to use the YouTube API to retrieve the corresponding channel ID, as well as metadata on the video, such as the video’s category, title, and duration, for all unique video IDs. From this metadata, researchers were able to then label each video based on the political leaning of its channel. Videos were categorized into “a single set of six categories: far left (fL), left (L), center (C), anti-woke (AW), right (R), and far right (fR)” (Hosseinmardi et al., 2021).

Upon conducting this study, the results showed that though YouTube is dominated by mainstream and centrist sources, there does exist a group of above-average engaged users who consume far-right content. A high correlation between on-platform consumption and off-platform consumption of far-right content was found, which shows that users with a certain preference direct traffic to far-right content more so than the effects of the recommendation algorithm itself. Links between the communities that consume “anti-woke” content and far-right content were also found, but there was no evidence that the anti-woke communities serve as a gateway to far-right communities. Overall, the findings suggest that YouTube serves more as a database of content that users approach with certain beliefs and intentions that reflect their general web browsing behavior. Researchers believe that though YouTube has an undoubtedly wide reach, it is still “just a part of an even larger information ecosystem that includes the entire web, along with TV and radio”, and thus, “the growing engagement with radical content on YouTube may simply reflect a more general trend driven by a

complicated combination of causes, both technological and sociological, that extend beyond the scope of the platform's algorithms and boundaries" (Hosseinmardi et al., 2021).

Though it appears that YouTube's current algorithm does not promote radicalism, a quick dive into its history before 2017 shows that this was not always the case. In 2017, YouTube along with many other global platforms like Facebook, Twitter, and Microsoft, created the "Global Internet Forum to Counter Terrorism" (Murthy, 2021). This was created in order for the platforms to become more accountable and take active measures towards combating extremist content online. As a result of these measures, extremist content has been found to have decreased substantially from these efforts, as evidenced in the research study by Ledwich & Zaitsev, finding YouTube not responsible for promoting radicalism. However, prior to this effort, it appears that YouTube's earlier algorithms may have been responsible for recommending extreme content. In the paper "Evaluating Platform Accountability: Terrorist Content on YouTube", researchers used 11 ISIS videos as seeds to build a network of 15,021 videos with 190,087 edges (Murthy, 2021). Researchers then used a qualitative comparative analysis to evaluate the eleven different video attributes to identify a key set of attributes found to potentially be involved in the recommendation algorithm recommending extreme content. By examining the network, which was built on videos and recommendations from 2016, researchers were able to make a few conclusions about YouTube's algorithm prior to 2017: First, official ISIS content was searchable and being incorporated within recommendations. Second, ISIS videos tended to recommend other ISIS videos. Third, some non-ISIS videos did recommend ISIS videos, albeit a rare occurrence and seemingly due to shared metadata attributes with ISIS videos (particularly having a similar title). After comparing the 2016 network to the YouTube network of 2020, it was found that ISIS videos were no longer being recommended, and that it was either extremely difficult or impossible to search for ISIS content.

In comparing the data on the amount of radicalizing or terrorist content present on YouTube's platform and the amount of content that is recommended by its algorithm, both before and after the formation of the Global Internet Forum to Counter Terrorism in 2016 and YouTube's pledge to decrease ISIS and terrorist content, research has shown that the amount of extremist content has decreased substantially, and that the recommendation algorithm actively discourages viewers from viewing this material. YouTube has shown that when active efforts are made to decrease and control dangerous content on its platform, namely, in the case of terrorist content, YouTube is successful in directing users to more mainstream content. It is important that the popular platform continues this trend, and proactively controls other dangerous extremist content that exists on the application.

Spread of Misinformation

YouTube's search and recommendation algorithms promote the spread of misinformation for a variety of controversial and potentially dangerous topics, all which have a potentially substantial impact. One such topic has dominated the past few years - information on the COVID-19 vaccine - and misinformation on the vaccine has proven to have wide-spread implications on the safety and stance of people's views on vaccines distributed against the current pandemic. Furthermore, COVID-19 vaccine misinformation is not just limited to the United States. As YouTube has seen consistent growth since its creation in 2005, the platform's user base has expanded globally, making its content extremely widely reachable and accessible. Moreover, as YouTube is becoming a platform that is used even in low and middle-income countries, the presence of health misinformation could lead to sustained pandemic outbreaks in areas of low immunization rates and access to vaccines.

Debates over vaccines and vaccinations are not new. Prior to the COVID-19 pandemic, as well as COVID-19 immunizations, there have existed anti-vaccination movements, which have gained prevalence over the last twenty years. The cause for anti-vaccination viewpoints can partly be attributed to the pervasiveness of vaccine misinformation that exists on the internet, and YouTube is a platform that has perpetuated the spread of vaccine misinformation, partially due to their recommendation system.

In 2016, researchers conducted a study to examine the interlinkages between pro-vaccine and anti-vaccine videos, in an effort to aid public health professionals explore new ways to reach anti-vaccine audiences. In their study, "Examining Sentiments and Popularity of Pro- and Anti-Vaccination Videos on YouTube", researchers collected 9,489 recommended videos using

keywords such as “vaccines”, and then manually identified 1,984 of the videos to be directly relating to vaccines. The authors then categorized this subset of videos to be either pro-vaccine, anti-vaccine, or neutral, depending on their vaccine sentiment. Following this categorization and an analysis of their results, researchers concluded that anti-vaccine videos not only made up the majority of the vaccine-related videos on YouTube at 65.02%, but that they also had “higher values of closeness centrality, suggesting that watching an anti-vaccine sentiment videos will likely lead to more anti-vaccine video recommendations” (Song & Gruzd, 2017). Furthermore, anti-vaccine videos were significantly more prevalent in the “News & Politics” and “People & Blogs” categories, while pro-vaccine videos were more common in the “Education” and “Science & Technology” categories. This study highlighted the need for YouTube to diversify its recommendation methods, especially in cases of vaccination misinformation, to help viewers break out of anti-vaccine filter bubbles.

In 2019, following the pushing of the COVID-19 debate onto the forefront of pressing and controversial issues prevalent around the world, the same researchers conducted a similar study to examine and contrast the results related to vaccine sentiment and misinformation found in 2016, before the COVID-19 pandemic, and 2019, after the COVID-19 vaccine had been developed in “Examining Algorithmic Biases in YouTube’s Recommendations of Vaccine Videos”.

Researchers retrieved 8,425 YouTube videos that were related to vaccination and discovered a network of recommended videos. One author, who had expertise in the public health industry, then individually watched all of the selected vaccine-related videos to categorize each video into either a pro-vaccine video, neutral, or anti-vaccine. In particular, pro-vaccine videos were in support of immunization, while anti-vaccine videos offered primarily attitudes of refusal or rejection towards vaccines. Neutral videos were neither supportive nor opposed to vaccines, such as news media that were presenting to the two sides of the debate.

In contrast to the study, “Examining Sentiments and Popularity of Pro- and Anti-Vaccination Videos on YouTube” by some of the same researchers, which was conducted in 2016 before COVID-19 was prevalent, the authors first found that on the YouTube platform itself, pro-vaccine videos were more prevalent that anti-vaccine videos, the opposite of the case just three years prior. Specifically, there were nearly three times more pro-vaccine videos than there were anti-vaccine videos. The authors noted that their “previous study conducted using the same query terms and the same data collection tool in 2016 yielded almost an opposite result showing the prevalence of anti-vaccine over pro-vaccine videos then” (Abul-Fottouh et al., 2020). Furthermore, researchers found that pro-vaccine videos were more likely to be recommended by YouTube than anti-vaccine videos were, which is consistent with the results found by Hussein et al. This suggests that “YouTube’s demonetization policy of harmful content and other changes to their recommender algorithm might have been effective in reducing the visibility of anti-vaccine videos” (Abul-Fottouh et al., 2020). However, while there was indeed a higher prevalence of pro-vaccine videos on YouTube, there was also evidence of the presence of a filter bubble effect, as well as evidence of homophily, which refers to the idea that users who have similar beliefs, behaviors, or habits tend to act in a certain way, and therefore, video recommendations would also be clustered around similar sentiments towards topics such as vaccination. Therefore, YouTube, as well as other technology companies, has a need to design their recommendation systems to account for the quality and credibility of video content, especially in cases about health information.

In addition, in the audited study, “Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube”, researchers found similar results. Amid growing concerns that YouTube’s search engine contributes to promoting and recommending misinformative content for search topics, researchers conducted an audit to verify these claims. Researchers studied how YouTube’s search and recommendation algorithm may be affected by user attributes, such as gender, age, geolocation, and watch history, and if these characteristics affected five different commonly misinformative topics. These topics included conspiracy theories involving the 9/11 tragedy, the chemtrail conspiracy theory, the idea that the earth is flat, the conspiracy surrounding the moon landing, and vaccine controversies, especially in light of the COVID-19 pandemic. Auditors conducted two different sets of experiments - the first being a Search audit, which involved studying how YouTube’s algorithm recommended and produced search results for brand new user accounts, all in which had no existing watch history. The other experiment was a Watch audit, in which researchers examined user accounts who had “built watch history by systematically watching either all promoting, neutral, or debunking videos of

potentially misinformative topics" (Hussein et al., 2020). Researchers created over 150 Google accounts, and collected 56,475 YouTube videos across the five misinformative topics, and all corresponding to videos present either in the search results, the Up-Next, and the Top 5 recommendations features of YouTube.

While auditors found that there was little evidence supporting that user's age, gender, and geolocation provided any significant role in amplifying misinformation both in YouTube's search results or recommendations for the brand newly created user accounts, they found that for existing accounts, a user's watch history had a significant effect on the amount of misinformative content present in a user's search results for all five controversial topics studied, and particularly when relating to vaccine controversies. Interestingly, researchers observed a filter bubble effect in recommendations for all topics *except* for vaccine controversies. That is to say, while watching promoting misinformative videos led to the recommendation of more similar videos of misinformation in the Up-Next and Top 5 video recommendations, for the topic of vaccines specifically, this filter bubble effect was not observed in recommended videos, though it still existed in 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" (Hussein et al., 2020). These results were consistent with those found by Abul-Fottouh et al., in their study conducted in 2019. These results, equivalently, suggest that YouTube is modifying its search and recommendation algorithm for certain misinformative topics such as vaccine controversies, and that these modifications have succeeded in aiding public sentiment to steer towards more scientifically correct information.

As previously mentioned, YouTube's search and recommendation algorithms promote the spread of misinformation and can lead to filter bubbles, which have wide-spread implications on the impact on perpetuating user's existing biases and personalized feeds. Though the existence of these filter bubbles has been proven, there also exists a need to study how easily one can get out of this filter bubble. In "An Audit of Misinformation Filter Bubbles on YouTube: Bubble Bursting and Recent Behavior Changes", researchers studied, once in a filter bubble of information, what it takes for a user to then get out of this filter bubble, and how the algorithm can create a more neutralized recommendation feed. This revealed information on not just the process of YouTube's personalization process, but also how to "improve the social, educational, or psychological strategies for building up resilience against misinformation" (Tomlein et al., 2021).

Researchers simulated user behavior on YouTube and recorded platform responses, both in the context of search results and recommendations provided by YouTube, and annotated these videos for the presence of misinformative content. They had predefined actions that were designed to first invoke the misinformation filter bubble effect by watching videos that contained misinformative content, and then seeked to mitigate the filter bubble effect by watching videos that consisted of trustworthy content. Auditors then were able to quantify the dynamics of the creation of the misinformation filter bubble, as well as the dynamics of the bursting of this bubble to reach a more balanced recommendation feed. In this process, researchers simulated fifty runs per day, each time watching 3,951 videos over the course of 50 stimulations, and recorded a total of 17,405 unique videos that were recommended to them.

Through this process, researchers found that though the bursting of a filter bubble is indeed possible, it is manifested differently depending on the topic. Consistent with the studies performed by Hussein et al., as well as Abul-Fottouh et al., the only topic to show a statistically significant difference and improvement was the topic regarding anti-vaccination, again, suggesting that YouTube's efforts to decrease the amount of misinformation regarding anti-vaccination following the COVID-19 pandemic has indeed decreased the filter-bubble effect on this topic alone. However, researchers also found that there were not many improvements in misinformation occurrences regarding other topics, despite YouTube's pledges to improve on the spread of misinformation through their platform. This calls for "the need for independent oversight of personalization behavior of large platforms" (Tomlein et al., 2021).

Conclusion and Future Directions

With the exponential increase in information on the internet and the use of social media, digital platforms have become a common source for people to get their news from. Platforms like YouTube have spent many years developing algorithms to better the experience on the platform. Evolving from simple algorithms that prioritize watch time to complex algorithms that personalize recommendations for each user, these algorithms help aid users in navigating and consuming information online. However, these algorithms also give powerful global platforms like YouTube the ability to not only influence political and extremist viewpoints, but also to increase the spread of misinformation. Because YouTube depends on viewers' retention time on videos and is ultimately economically driven, they have little incentive to stop recommendations that may be producing politically polarized communities. The personalization of the content and users' inherent desire to seek information that aligns with their own creates an environment on the platform that fosters echo chambers and filter bubbles that reinforce users' pre-existing beliefs, ultimately creating a larger divide in the political community. Research has shown that recommending content based on a user's social circle could potentially minimize the effects of selective exposure and encourage interaction with differing viewpoints. However, this phenomenon is not only due to YouTube's algorithm, but can be traced back to other factors like filter bubbles and echo chambers that work together with cognitive factors, such as confirmation bias, that also increase polarized communities in our society. Through our research, we have also found that the actual content promoted on platforms is a key factor in producing homogeneous echo chambers regardless of the algorithm. The increase in political polarization in our society may be a larger issue that expands beyond the scope of just the YouTube platform. When it comes to conspiracy theories and rabbit holes on YouTube, it has generally been found that visitors to the site are recommended these types of videos despite having no interaction with such content beforehand. Several studies supported these findings through different methods that took advantage of data crawlers and co-visitation counts. Especially when considering the concept of technological seduction used to create recommender systems and the natural tendency of humans toward selective exposure, it is clear that the way YouTube categorizes and associates similar videos is flawed and potentially dangerous in leading to extremist viewpoints and rabbit holes.

In terms of extremist content and the spread of misinformation regarding COVID-19 vaccine information, our research has shown that when YouTube has made active efforts to decrease the amount of potentially dangerous videos that promote radical content, as they did in 2017, or when YouTube made a pledge to flag and decrease the amount of COVID-19 vaccine misinformation, public sentiment was steered towards more scientifically correct information, and the amount of terrorism-promoting content has decreased significantly. Both of these outcomes came after there was public outcry for YouTube to be accountable regarding these topics. In 2017, global platforms, including YouTube, Facebook, Twitter, and Microsoft, created the "Global Internet Forum to Counter Terrorism" following an international movement to counteract terrorism. In 2020, following the global movement to promote COVID-19 vaccinations and immunizations, research has shown that the amount of vaccine misinformation present on YouTube has decreased substantially, compared to just three years prior. However, there still exists a plethora of misinformation, misleading content, and potentially dangerous content that exists on the platform today. Therefore, YouTube's recommendation system has a need to be reformed to become proactive instead of reactive, and more effort needs to be made to control the spread of other misinformation and extremist content, beyond what is just publicly called for.

Moving forward, we would want to look into the broader implications of digital media and its effects on user behaviors, and what YouTube specifically can do to mitigate the harmful effects of filter bubbles and echo chambers that emerge from its platform. Specifically we would want to dive deeper into the causes of the demassification of mass communication due to the wide spreading self-selectivity encouraged by the increase of digital media and online news sources and look at other platforms that have similar harmful effects. We have also seen that when YouTube makes an active effort to decrease the recommended content from potentially dangerous and misleading information, their efforts have succeeded, and have been effective in steering public opinion towards more neutral viewpoints due to their large outreach and influence. However, further research can be done on how exactly YouTube succeeded in decreasing this harmful content on their platform, so that future efforts can be made to decrease other misleading and dangerous content. Whether these efforts were made with recommendation algorithm remediations, more live and human moderation instead of relying

heavily on an algorithm to control content, better marketing strategies, or other strategies, remediations have proven to be successful, and therefore it is important to study the specifics of how YouTube was able to decrease the amount of harmful and misleading content. Moreover, researchers can use this information to develop better marketing strategies, design more accountable recommendation systems, and use their platforms to promote positive and uplifting content.

Sources

Abul-Fottouh, D., Song, M. Y., & Gruzd, A. (2020). Examining algorithmic biases in YouTube's recommendations of Vaccine videos. International Journal of Medical Informatics, 140, 104175. https://doi.org/10.1016/j.ijmedinf.2020.104175

Alfano, M., Fard, A. E., Carter, J. A., Clutton, P., & Klein, C. (2020). Technologically scaffolded atypical cognition: the case of YouTube's recommender system. Synthese, 199(1-2), 835–858. https://doi.org/10.1007/s11229-020-02724-x

Bessi, A., Zollo, F., Del Vicario, M., Puliga, M., Scala, A., Caldarelli, G., Uzzi, B., & Quattrociocchi, W. (2016). Users polarization on Facebook and YouTube. *PLOS ONE*, *11*(8). https://doi.org/10.1371/journal.pone.0159641

Cho, J., Ahmed, S., Hilbert, M., Liu, B., & Luu, J. (2020). Do search algorithms endanger democracy? an experimental investigation of algorithm effects on political polarization. *Journal of Broadcasting & Electronic Media*, *64*(2), 150–172. https://doi.org/10.1080/08838151.2020.1757365

Cohen, J. N. (2018). Exploring echo-systems: How algorithms shape immersive media environments. Journal of Media Literacy Education, 10(2), 139–151. https://doi.org/10.23860/jmle-2018-10-2-8

Covington, Paul, et al. "Deep Neural Networks for YouTube Recommendations." Proceedings of the 10th ACM Conference on Recommender Systems, 2016, https://doi.org/10.1145/2959100.2959190.

Feezell, J. T., Wagner, J. K., & Conroy, M. (2021). Exploring the effects of algorithm-driven news sources on political behavior and polarization. *Computers in Human Behavior*, *116*, 106626. https://doi.org/10.1016/j.chb.2020.106626

García-Marín, J. (2021). YouTube and traditional media: Polarization in the Catalan Political Conflict. *Studies in Digital Politics and Governance*, 31–41. https://doi.org/10.1007/978-3-030-71815-2_3

Hosseinmardi, H., Ghasemian, A., Clauset, A., Mobius, M., Rothschild, D. M., & Watts, D. J. (2021). Examining the consumption of radical content on YouTube. *Proceedings of the National Academy of Sciences*, *118*(32). https://doi.org/10.1073/pnas.2101967118

Hussein, E., Juneja, P., & Mitra, T. (2020). Measuring misinformation in video search platforms: An audit study on YouTube. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW1), 1–27. https://doi.org/10.1145/3392854

Landrum, A. R., Olshansky, A., & Richards, O. (2021). Differential susceptibility to misleading flat earth arguments on YouTube. *Media Psychology, 24*(1), 136-165). https://doi.org/10.1080/15213269.2019.1669461

Ledwich, M., & Zaitsev, A. (2020). Algorithmic extremism: Examining YouTube's Rabbit Hole of Radicalization. *First Monday*. https://doi.org/10.5210/fm.v25i3.10419

Murthy, D. (2021). Evaluating platform accountability: Terrorist content on YouTube. *American Behavioral Scientist*, *65*(6), 800–824. https://doi.org/10.1177/0002764221989774

O'Callaghan, Derek, et al. "Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems." Social Science Computer Review, vol. 33, no. 4, 2014, pp. 459–478., https://doi.org/10.1177/0894439314555329.

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A., & Meira, W. (2020). Auditing radicalization pathways on YouTube. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3351095.3372879

Song, M. Y.-J., & Gruzd, A. (2017). Examining Sentiments and Popularity of Pro- and Anti-Vaccination Videos on YouTube. Proceedings of the 8th International Conference on Social Media & Society - #SMSociety17. https://doi.org/10.1145/3097286.3097303

Tang, L., Fujimoto, K., Muhammad (Tuan) Amith, Cunningham, R., Costantini, R. A., York, F., . . . Tao, C. (2021). "Down the rabbit hole" of vaccine misinformation on YouTube: Network exposure study. *Journal of Medical Internet Research, 23*(1). http://dx.doi.org/10.2196/23262

Tomlein, M., Pecher, B., Simko, J., Srba, I., Moro, R., Stefancova, E., Kompan, M., Hrckova, A., Podrouzek, J., & Bielikova, M. (2021). An audit of misinformation filter bubbles on YouTube: Bubble bursting and recent behavior changes. Fifteenth ACM Conference on Recommender Systems. https://doi.org/10.1145/3460231.3474241