The Danger in YouTube’s Algorithm and How We Can Prevent It

Carissa Kieger with Professors Chun and Elkins
IPHS 200 Programming Humanity - Fall 2021

Introduction

YouTube, founded in 2005, is currently the strongest video-hosting platform in the world with over 2 billion users. Hundreds of hours of YouTube content are uploaded every minute, and over a billion hours of video content are consumed by users every single day. Since a significant portion of content that viewers watch is brought to them by YouTube’s recommendation feature, it is important to analyze and critique it to ensure that it performs at the highest standard possible. This project will explore the flaws within YouTube’s recommendation system, what actions YouTube has taken to address them, and where improvements still need to be made.

While YouTube maintains a constant presence in video streaming, it has struggled to keep a safe platform free from negative biases and features that taint its recommendation system. By putting three studies discussing biases, radicalization, and reliability into conversation with each other as well as YouTube’s report, we will determine what the next steps need to be in order to for YouTube to maintain a safe platform for reliable information and entertainment.

Background

In 2008, YouTube started its first recommendation system, which ranked uploaded videos based on view count. Since then, YouTube’s recommendation system has evolved into a machine learning system that incorporates user viewing and search history, watch time, and content engagement in a “black-box” algorithm. With little visibility of its inner operations, data is collected and inputted into said algorithm, and proceeds to output personalized recommendations for individual users. However, using this method to promote certain content poses many concerns, such as those pertaining to ethics, privacy, and accuracy.

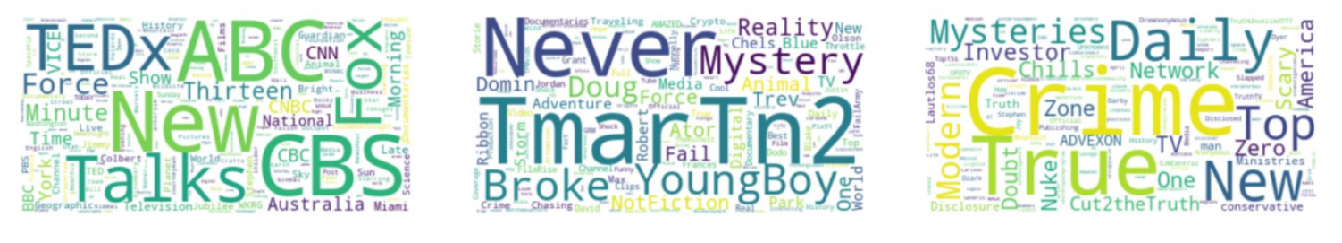
There has been a great deal of research on the specifics of YouTube’s recommendation system, and although it has proven to be quite successful at engaging users, flaws have been discovered as well. The algorithm that runs the suggestions feature has exploited societal biases– for example, LGBTQ+-related videos in the past have been labeled as “potentially offensive content.” Research has also indicated that since watch history is heavily incorporated into the suggested videos algorithm, filter bubbles form, isolating the user from any content that might challenge their values or beliefs. Both of these instances indicate that there are issues within the algorithm that lead to the formation of echo-chambers.

In 2017, after specific concerns were raised of the algorithm holding biases against marginalized communities, YouTube took a pledge to analyze its machine-learning algorithm to ensure fair treatment across all groups. Additionally, YouTube began favoring “authoritative” content within its recommendation system.

YouTube’s recommendation system gained traction in media following the 2016 U.S. Presidential Election, with media outlets such as the New York Times criticizing YouTube for its “radicalizing” power. In 2019, YouTube formulated a response with a promise to demote content deemed “borderline” to its community guidelines.

Methodology & Results

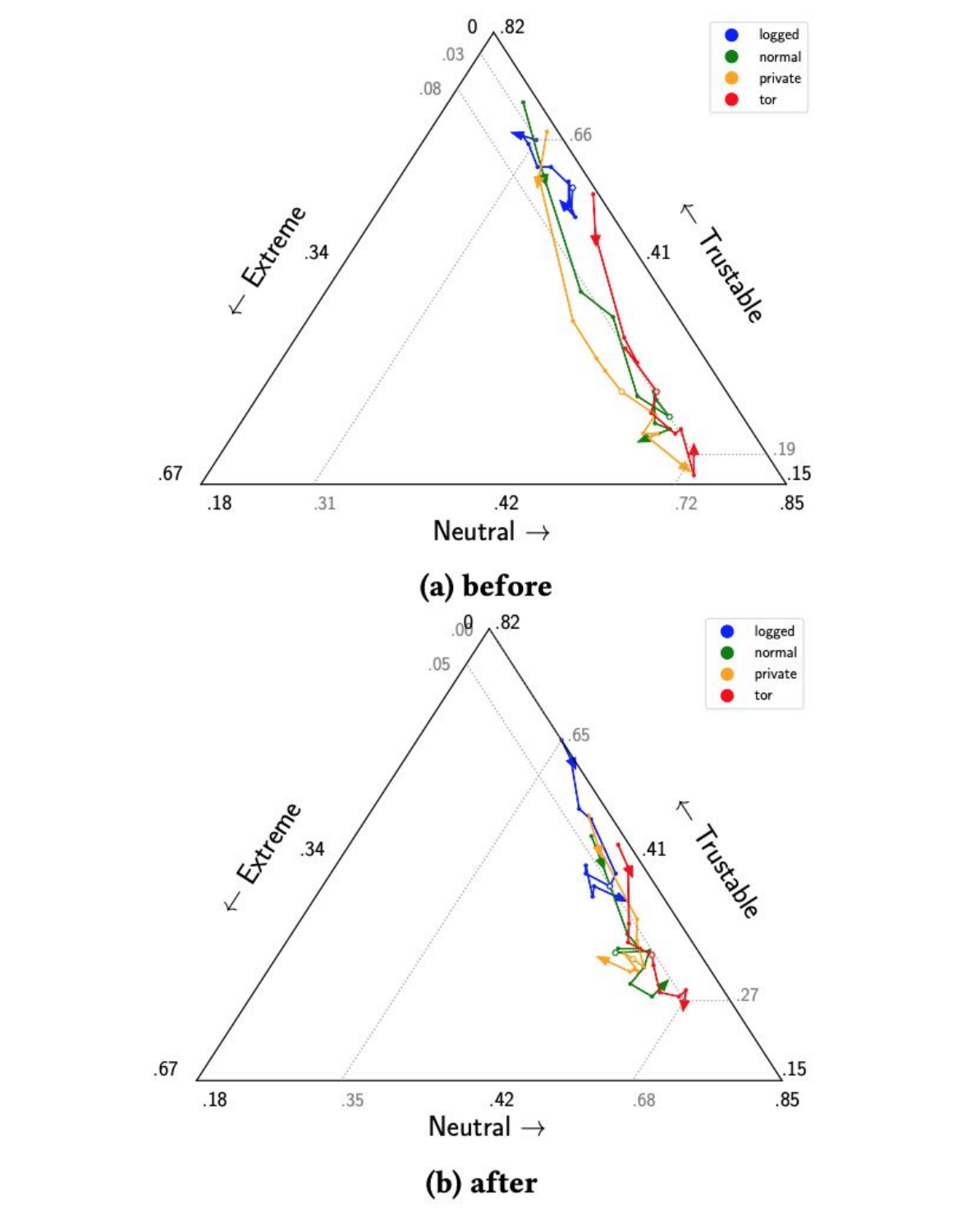
Prior to YouTube’s largest recommendation announcement stating that content would be demoted, a study run by Spinelli & Crovella had begun on its algorithmic system. The framework used for data collection was to track “chains” of recommended content by YouTube for analysis on the suggested videos over time. The study began by utilizing four privacy scenarios (logged, normal, private, and Tor) to determine if the way a user watches videos has any effect on what content is recommended to them. To label the suggested videos, they were evaluated within three categories: “Trustable,” “Neutral,” and “Extreme” content. Below is an example of the word cloud classifications.



(a) Trustable (b) Neutral (c) Extreme

Figure 1: Channel Name Word Clouds by Classification.

Each video in the suggestion chain was classified within the range of reliability and plotted on a ternary plot. Since Spinelli & Crovella’s study began before the 2019 policy change and ended after it had been implemented, they were able to track the initial effects of the algorithm adjustment on “extreme” content recommendations to user instances, and whether or not it was as effective as YouTube claimed it to be. Displayed to the left is a data visualization tracking user access method and recommendation shifts over time.



(a) before

(b) after

Figure 11: Recommendation Shift Before and After YouTube Policy Change, by Privacy Scenario.

Masadeh & Hamilton’s study focused primarily on biases within YouTube’s recommendation system and occurred after YouTube’s policy change in 2019. The study states that many creators claim to experience negative effects following the change, reporting that their videos are no longer recommended to users that are not already subscribed to their channels. To test this claim, Masadeh & Hamilton searched different political topics on their personal YouTube account to see the recommended results.

Since YouTube’s business model relies heavily on maintaining user engagement for advertisements, the autoplay feature selects the top-recommended video according to the algorithm for the viewer to watch following the video they’re on. Markmann & Grimme conducted a study that tracked autoplay recommendations in order to explore the inner workings of the algorithmic system. Their study tracked the progression of the autoplay feature and frequency of channels on recommended videos. By allowing autoplay to run on 30 accounts for ten videos each, Markmann & Grimme compared their findings for frequency of video similarity for short and long videos across the subjects of “News”, “Music”, “Trends”, and “COVID-19.” Their findings are provided in the following table (labeled Table 2).

Results & Analysis

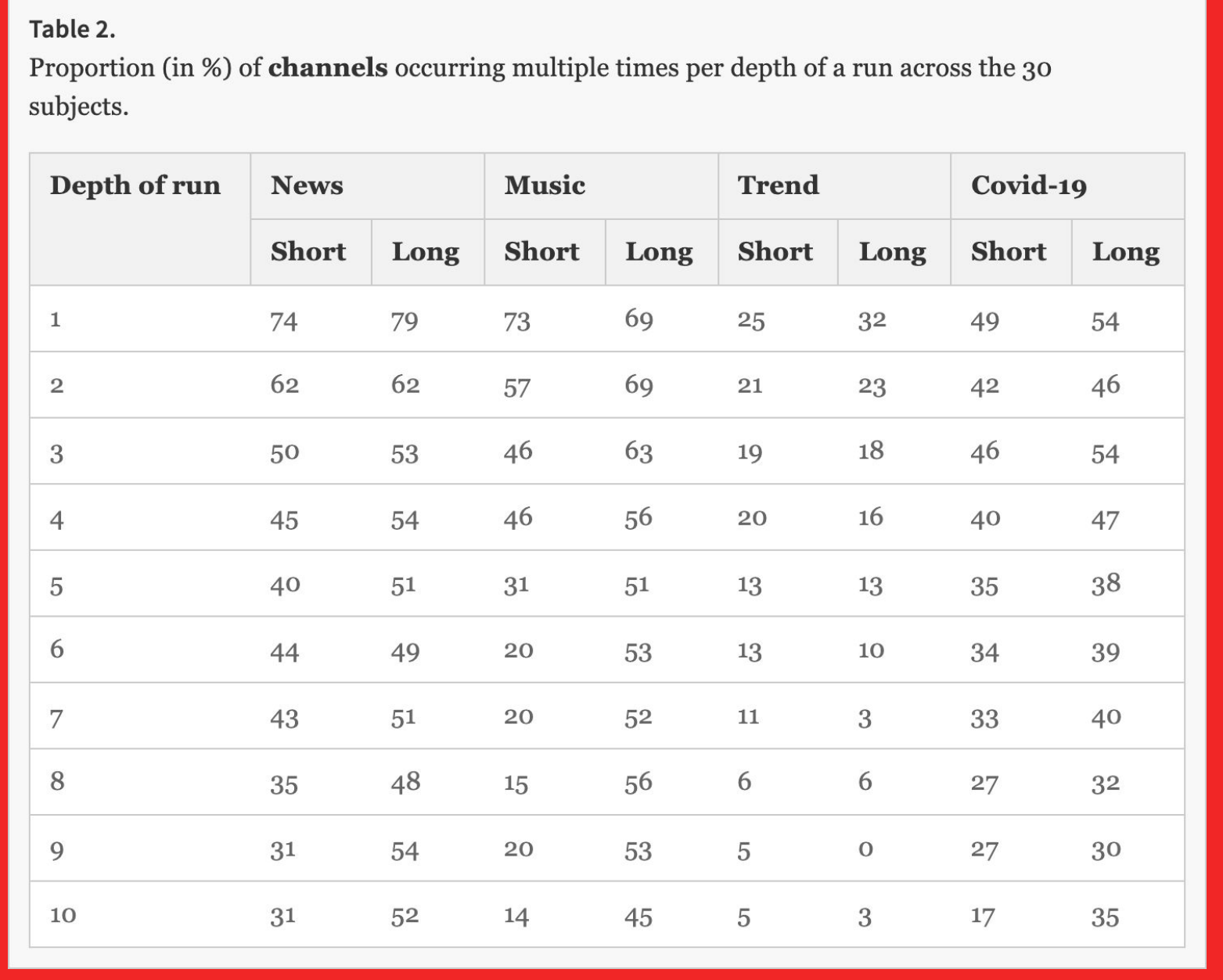


Table 2.
Proportion (in %) of **channels** occurring multiple times per depth of a run across the 30 subjects.

Depth of run	News		Music		Trend		Covid-19	
	Short	Long	Short	Long	Short	Long	Short	Long
1	74	79	73	69	25	32	49	54
2	62	62	57	69	21	23	42	46
3	50	53	46	63	19	18	46	54
4	45	54	46	56	20	16	40	47
5	40	51	31	51	13	13	35	38
6	44	49	20	53	13	10	34	39
7	43	51	20	52	11	3	33	40
8	35	48	15	56	6	6	27	32
9	31	54	20	53	5	0	27	30
10	31	52	14	45	5	3	17	35

According to Markmann & Grimme’s findings, YouTube’s autoplay and recommendation feature produce groupings of frequently recommended news channels. While more information-based categories such as “News” and “COVID-19” have a high cosine similarity of 0.652, “News” in comparison to “Trend” has a cosine similarity of 0.043. They conclude that this occurs because there are only a select number of channels that are recommended within the news category, whereas viral videos enter the “Trend” category regardless of the channel name.

These findings are echoed in Masadeh & Hamilton’s study on YouTube’s systemic bias. In pulling information on recommended videos from search results, Masadeh & Hamilton noted that although the account was subscribed to over 15 political and news channels not considered within the “mainstream media” realm, none of the top results featured those channels, and all of the suggestions were from corporate news outlets. These results, along with Markmann & Grimme’s findings, confirm the 2017 evaluation and decision by YouTube to favor “authoritative” content in its recommendations, which the authors claim reinforce different cognitive biases. An unforeseen consequence of this implementation has been a reliance on larger channels that hold a presence across platforms, with smaller content creators at a disadvantage due to lack of “credibility.”

The specific media outlets that are highly prioritized within the recommendation system are not directly addressed within Spinelli & Crovella’s research. However, the limitations portion of the study does briefly discuss concerns over their definition of the categories “Trustable,” “Neutral,” and “Extreme.” Their findings of most frequent channels within the “Trustable” category (e.g. Fox, CBS, TEDx) support Masadeh & Hamilton as well as Markmann & Grimme’s conclusions that “authoritative” content within YouTube’s recommendation system is driven by bias towards a select few corporate and mainstream media outlets.

According to YouTube, there was a 70% decrease in watch time on “borderline” content in the U.S. that stemmed from recommendations. However, Spinelli & Crovella’s findings, illustrated in “Figure 11” signify that although there is less of a change from “Trustable” to “Extreme” content as the recommendation chain progresses, the shift towards “radical” content is still very prevalent. Prior to the change, the path had an 8.5% increase in the fraction towards “Extreme” recommendations, whereas, after the change, the path had a 5.9% increase. These numbers signify that despite YouTube’s report that watch time on “borderline” content has decreased significantly, the content continues to be promoted and recommended to users over time.

Conclusion

As Goodrow, YouTube’s VP of Engineering, states, “Misinformation… tends to lack a clear consensus, and can vary depending on personal perspective and background… sometimes, this means leaving up controversial or even offensive content.” While YouTube has taken a stance that it will only remove content that clearly violates its community guidelines to avoid censoring individuals, it continues to suppress smaller content creators in favor of “Authoritative” mainstream media.

For YouTube to follow in accordance with its statement, I propose that the promotional feature of “Authoritative” content is removed. In its place, in-depth channel descriptions would be provided on every channel that holds a “verified” checkmark. Built into YouTube’s user interface, the description would be public to all users and would provide an objective and comprehensive summary of the channel’s background. By including factual informative rather than persuasive information on channels, users will be fully informed on the channel’s background and possible motivation for content production.

Transparency is key to removing misinformation and avoiding censorship. With added contextual information, both issues discussed in this project of “Authoritative” and “Extreme” content drowning out all others would be mitigated. Although there is currently no solution to YouTube’s algorithm pushing more “radical” content since it utilizes a black-box algorithm, the context provided on such videos would inform users as to what ways the video might be using ulterior motives to achieve certain goals. From there, power would be placed in the user’s hands in terms of what information they choose to rely on.

In order for us to be able to trust YouTube to produce safe and accurate recommendations, we must continue to experiment on the recommendation algorithm to expose its biases and inequities. Overall, there has been an extensive amount of research on YouTube’s recommendation system, however, it continues to be put out of date due to its machine-learning capabilities and YouTube’s reevaluations that lead to adjustments in its algorithm. By continuously researching and critiquing the system, we may track YouTube’s algorithm advancements and adaptability. In implementing the proposed feature that provides channel context, further data may be collected to determine whether or not this proves to be successful. In order for changes to occur, the responsibility lies on users (who experience the system firsthand), to expose flaws within YouTube’s system. Our research and feedback are the main components that will hold YouTube and its algorithm accountable.

References

- Goodrow, Cristos. (2021). On YouTube’s Recommendation System. *YouTube Official Blog*. https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/
- Markmann, S., & Grimme, C. (2021). Is Youtube Still A Radicalizer? An Exploratory Study On Autoplay And Recommendation. *Disinformation In Open Online Media*, 50-65. doi:10.1007/978-3-030-87031-7_4.
- Masadeh, S., & Hamilton, B. (2020). The Aftermath Of The Adpocalypse: Systemic Bias on YouTube. *CHI*. http://critical-media.org/cobi/docs/Masadeh_Bias%20on%20YouTube.pdf.
- Spinelli, L., & Crovella, M. (2020). How Youtube Leads Privacy-Seeking Users Away From Reliable Information. *FairUMAP*, 244-251. https://dl.acm.org/doi/10.1145/3386392.3399566.
- Suciu, P. (2021). Youtube Remains The Most Dominant Social Media Platform. *Forbes*. https://www.forbes.com/sites/petersuciu/2021/04/07/youtube-remains-the-most-dominant-social-media-platform/?sh=5a0f8c456322.
- YouTube. (2022). *Brand resources - How YouTube Works*.https://www.youtube.com/howyoutubeworks/resources/brand-resources/#logos-icons-and-colors