Understanding the Radicalizing Effects of Recommendation Algorithms

Sasha Rabeno
Dept. of Industrial Systems & Engineering, Lehigh University
Bethlehem, Pennsylania
aer224@lehigh.edu

***Abstract*—Algorithms that recommend personalized content to users are a staple of nearly all social media platforms. However, these algorithms often push users towards ideologically extreme content, with the potential to escalate from on-screen hate to off-screen violence. I created a scoring function to create a list of videos for a given user to watch that best align with their preferences. This research finds that with our scoring system in place, users are pushed to watch videos more extreme than they would in the absence of a recommendation system.**

Keywords—YouTube, recommendation algorithms, online extremism

I. How Did We Get Here?

So, what was the impetus for this project? The answer is I spend far too much time online, particularly on YouTube. I somehow went down the rabbit hole of online radicalization/echo chambers/extremism, and became interested in how social media companies are liable for creating extremists.

II. Project Planning

Thanks to funding from the Clare Booth Luce Scholarship Program, I was able to start this project in late May of 2023. The planning phase was mostly doing lots of reading. Papers read mainly covered political extremism, online echo chambers, and recommendation algorithms. There isn't much scholarship looking into the overlap of these areas (algorithms that push people towards political extremes), but the literature that does exist has conflicting stances. I began collecting figures from the papers I read, mostly about the political and psychological aspects of my project---how do people use YouTube? What do they use it for? How much politically extreme content is there really on YouTube?

My inspiration for the project's format came from an entirely unrelated YouTube video modeling how the Disney amusement park FastPass ticket system worked [2]. This video utilized an agent-based simulation to model how large numbers of guests interacted with different features of a park, neatly encapsulated in objects and classes. The source code for this simulation was online, so I spent time looking at how the simulation was organized. My initial drafts were a much-more pared down version of this simulation.

Once I decided upon having a class-based system, I outlined what attributes and objects my system would have. People and videos were the only two objects I found necessary, and each has a similar set of attributes. People (or "agents") have the following qualities:

longest_vid_threshold: the longest video length (in minutes) a user will watch

yt_time_threshold: how much time (in minutes) a user spends on my made-up YouTube per day

political_affiliation: either "left," "right," or "middle" in the random system, but more specific in the archetype-based system---more on that later

video_extremity: the extremeness of videos that the user prefers, on a scale from 0.0 – 1.0

popularity_threshold: the preferred amount of views a video will have; a minimum

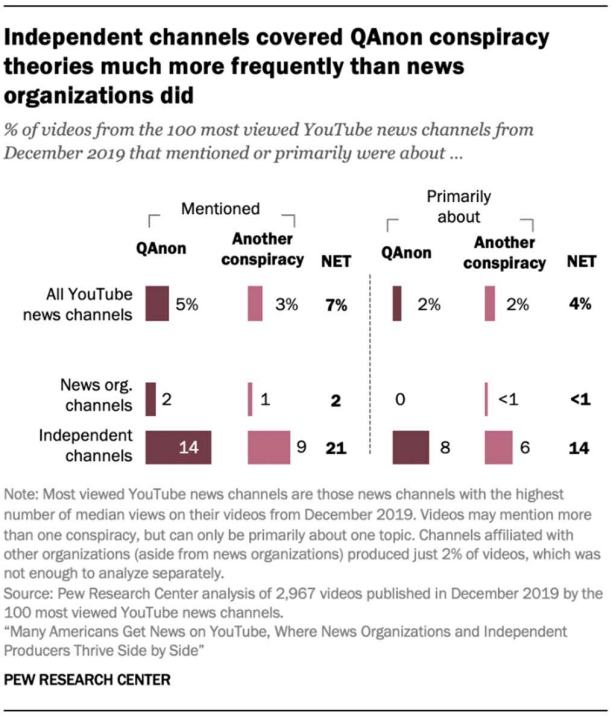


Fig. 1: One of the figures I referenced during the reading phase. This one is from the Pew Research Center [1].

Videos have a similar set of attributes, as follows:

views: exactly what it sounds like---number of views

vid_id: an integer uniquely identifying each video

length: how long a video is, in minutes

extremeness: how politically extreme a video is (0.0 – 1.0)

thumbs_up: wound up not being used, but the number of thumbs-up votes a video has

I spent a lot of time figuring out how I was going to define the people for this simulation. For videos, I ballparked what I figured a reasonable range of video lengths was (e.g. all videos are from 1 to 90 minutes in length), but how can I make those generalizations about users when their behavior is a critical part of the simulation? I wound up heavily referencing a study I found called Hidden Tribes [3] that claimed to break up the American population into seven "archetypes" of political thought. These archetypes encompass beliefs from all across the left-right political spectrum. In particular, the breakdown of what percentage of Americans made up each group directly translated to the percentage of each archetype users I had in the simulation. For example, 8% of my users were "Progressive Activists" (who preferred long, less-popular, very left-wing videos). 6% were "Devoted Conservatives," with the same preferences for length and popularity but with interest in very right-wing videos. Obviously, even this is a very simplified version of what actually happens on YouTube, but for the sake of this simulation it was good enough.

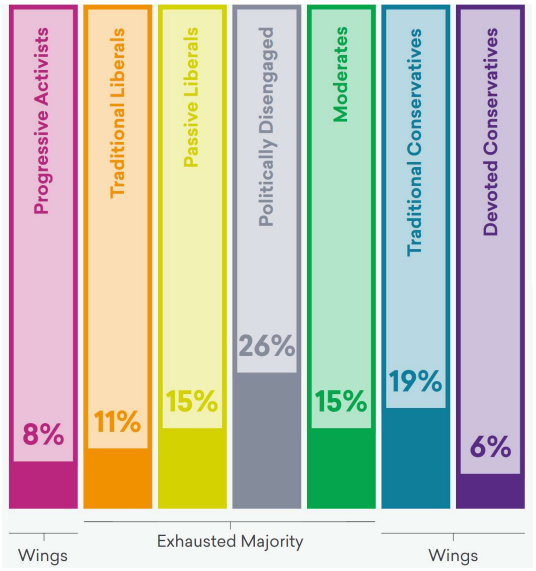


Fig. 2: The breakdown of Americans into seven political groups, in percentage of the U.S. population each group makes up [3].

III. Results with Archetypes

What did the simulation results look like with these archetypes in place? At this stage of the project, I was producing graphs for four different recommendation systems: a recommendation system (unhelpfully titled in graphs as the "rec system") that provides users a filtered list of videos to watch based on their preferences, a "random" system that's the same as the previous system but with a 10% chance of the user being shown a completely random video (think of all the random things YouTube or Facebook will sometimes show you), the more-specific "scoring" system, and no system at all.

My main focus of this simulation has been understanding how all these recommendation systems influence the extremeness of the content users watch. The following graphs plot a user's extremeness (from 0.0 – 1.0) against the average video extremeness they watched in a day.

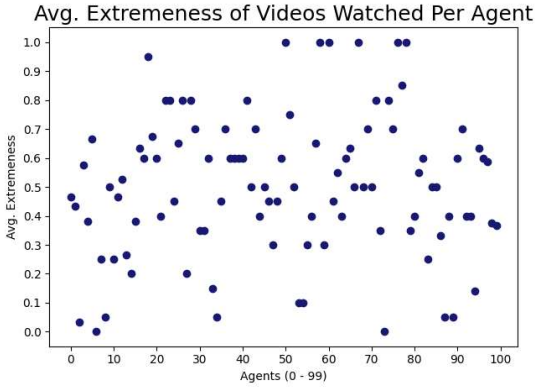


Fig. 3: Extremeness watched with no system.

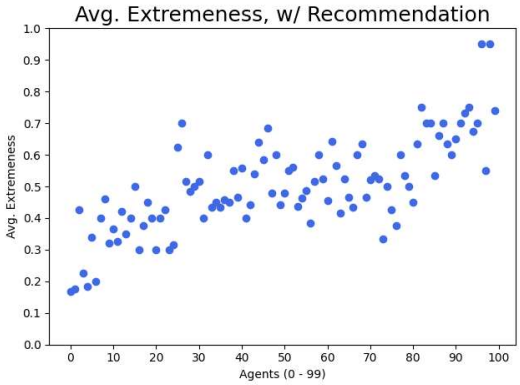


Fig. 4: Extremeness watched with filtration system.

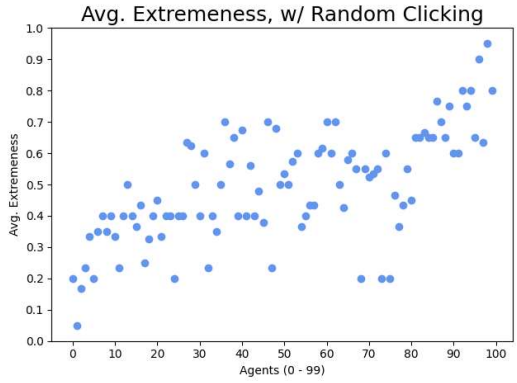


Fig. 5: Extremeness watched with filtration+ random clicking system.

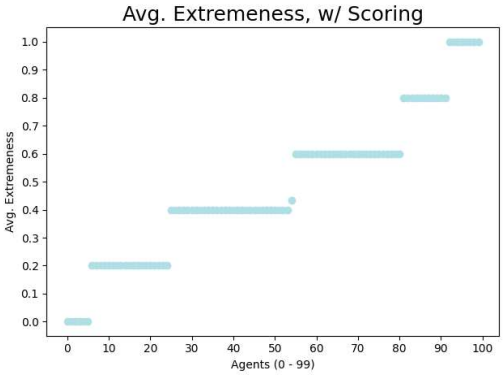


Fig. 6: Extremeness watched with scoring system.

The following figure is a regression plot of the previous four graphs. How convenient!

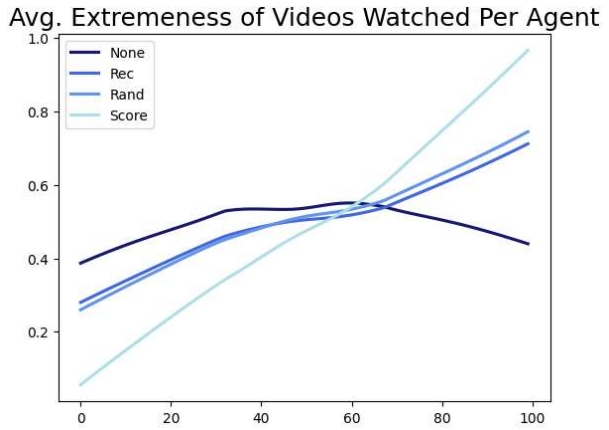


Fig. 7: Regression plots of the average extremeness watched for the same agents, with all four systems in place.

As we can tell from both this regression graph and the scatter plots, having a filtration system in place ("rec" and "rand" on the regression graph) generally leads users to watch content in their ideological wheelhouse. This isn't exactly a y=x type of curve, with a 1:1 relationship between a user's preferred extremeness and the content they actually watch, but the relationship is generally there. With no system, everybody watches everything. The most profound effects are with the scoring system, which offers a much more precise version of the filtration system. Here, users are pushed to watch much more extreme content than in the milder systems. This is a trend that will continue with further graphs. For now, let's look at the other graphs from this chunk of the project.

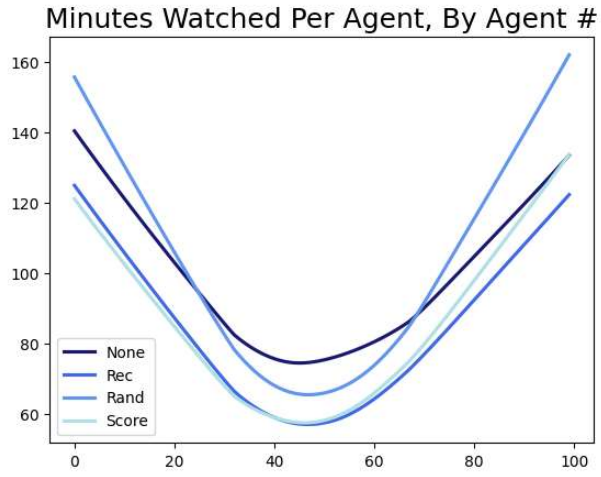


Fig. 8: The average amount of minutes watched per user, by user (they are sorted by extremeness on the x-axis).

This graph, Fig. 8, isn't anything too remarkable. The more "fringe" archetypes (closer to 0.0 and to 1.0) will watch longer videos and spend more time on YouTube. Thus, it makes sense that they are watching more, longer videos than the less politically engaged users. This relationship of less extreme agents preferring shorter videos (and for less time) continues in the next graph, Fig. 9.

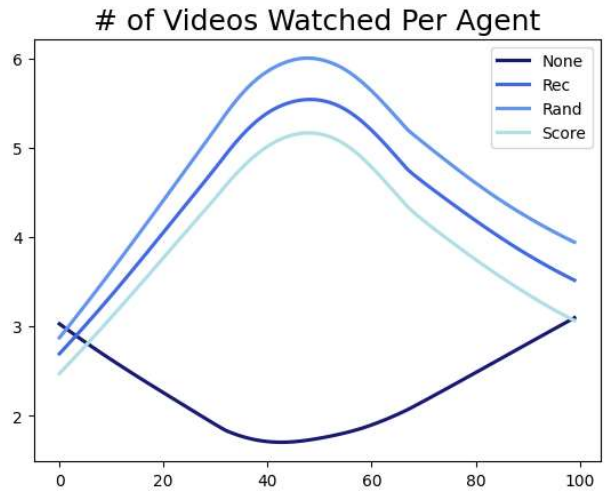


Fig. 9: Regression plots of the average number of videos watched for the same agents, with all four systems in place

Once again, this graph makes sense. With the systems in place, users behave according to their archetypes. The more extreme archetypes watch less videos in a day (because they're watching longer videos), while the middle-ground archetypes watch more videos, albeit much shorter ones.

I recreated these above graphs (Figs. 3 – 9) with just the middle three archetypes of agents (the middle 56% of users), to see if these same trends continued. As I anticipated, these same trends were present for even the middle agents.

IV. The Scoring System

The scoring system was the main focus of my work during the spring semester. Google has released some information about how their video recommendation algorithm works, but they're understandably tight-lipped about it. My main reference for how Google operates came from the horse's mouth: the short paper "Deep Neural Networks for YouTube Recommendations" published by Google. YouTube uses a deep learning system built on TensorFlow. Google claims that they "typically use hundreds of features in [their] ranking models," and that "our models learn approximately one billion parameters and are trained on hundreds of billions of examples" [5]. Their explanation of these "features" and parameters are incredibly vague. Because of this and my previous (unpleasant) experience with TensorFlow, we settled on mimicking the ranking model of YouTube via a scoring function. This scoring function would, ideally, generate a score for each user and video indicating how good of a match the two are. However, we could tweak this scoring system to mimic the factors we believe YouTube (and users) would find important. Thus, the scoring equation used is as follows:

$$\textbf{score} = -(\alpha \cdot \textbf{l\_return}) - (\beta \cdot \textbf{p\_return}) + \gamma * \textbf{abs(ev-ea)} - (\delta \cdot \textbf{e\_return})$$

Where:

α = weight placed on video length *(0.20)*

β = weight placed on video popularity *(0.15)*

γ = weight placed on video alignment (how similar video extremeness is to the user's) *(0.5)*

δ = weight placed on rewarding extremeness in either direction *(0.15)*

l_return will return 0 if the video length is greater than the user's preference, and 1 if it is less. This rewards shorter videos, as we assume that users do not want to spend lots of time on one video.

p_return will return 0 if the video's number of views is less than the user's preference, and 1 if it is greater. This rewards videos with a larger number of views.

abs(ev-ea) records the difference in magnitude between the user's extremeness (*ea*) and that of the video (*ev*). Videos that minimize this difference are rewarded, as we assume users want to watch videos that align with their ideological beliefs.

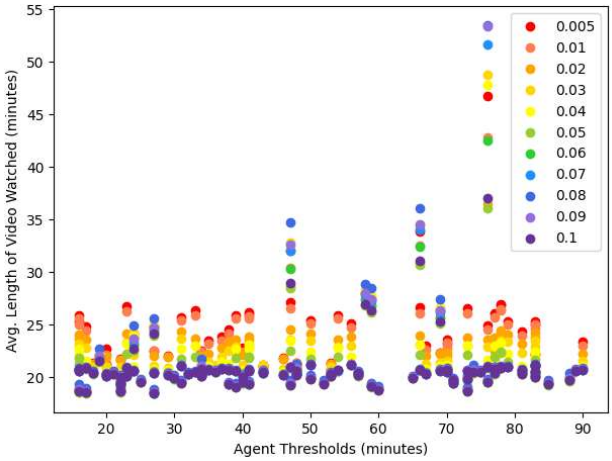
e_return will return 0.625 if the video has an extremeness ≤ 0.2 or ≥ 0.8, and 0 otherwise. This rewards videos that are more extreme.

This is a lot! To summarize: shorter videos are rewarded; more extreme videos (the top and bottom 20% of videos, by extremeness) are rewarded; videos that align with a user's extremeness preferences are rewarded; and videos that are popular (higher number of views) are rewarded.

At this point, we also switched to using users with randomly-generated preferences, rather than with archetypes. This presented similar results as with archetypes, which certainly made me feel better about all this data. These graphs also look a lot fancier. Here, we also decided to focus solely on the scoring system; the "rec" and "rand" systems are merely simpler versions of the scoring system, and we're trying to get closer to what it is that YouTube does.

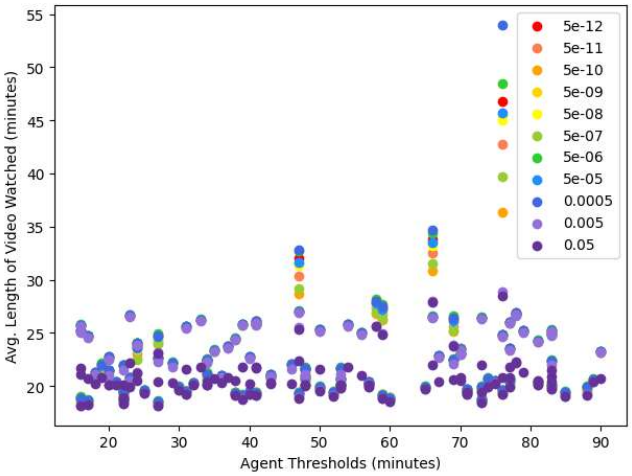
V. Double-Checking that the Weights on the Scoring System Are Accurate

The numerical weights assigned to each variable of the scoring function were decided pretty early on, but how could we know if these were the right values to be using? Enter about sixteen graphs' worth of testing. I spent a lot of time seeing how changing the weight on each variable of the scoring system affects all the other data points. For example, how does changing the weight on video popularity affect the popularity of videos users actually watch? Does video popularity affect extremeness watched? Not all of these graphs showed that much, but here are some of the highlights. If you wish to see just how many graphs I made, feel free to peruse the PowerPoint slides located in this project's GitHub repository [5].



This graph shows how, depending on the emphasis on alpha (video length), users watch shorter and shorter videos. This makes sense---a higher alpha value means videos are encouraged to be shorter and shorter. But why all the random scattered dots?

After meticulously checking how video scores vary with alpha, videos with longer lengths were shown to still get through the scoring system, even with a higher emphasis on videos being short. Thus, it makes sense that agents who can (those with higher thresholds) would be watching those videos (why the dots go up as the graph moves to the right), albeit not very much.



Here is the same graph, but with much smaller versions of alpha. I wanted to see how close to zero we could get without a weird stratification forming when alpha = 0, and it seems really close is the answer. We only see purple because those dots are covering the even smaller dots; we've hit a point where alpha is so small, it's no longer making much difference.

VI. Graphs for Randomly-Generated Users (AKA: The Graphs That Matter Here)

The graphs here are now green! Partially to differentiate them from the archetype graphs, but more so because my research poster was green. I really like the split in the data you can see between Figs. 10 and 11.

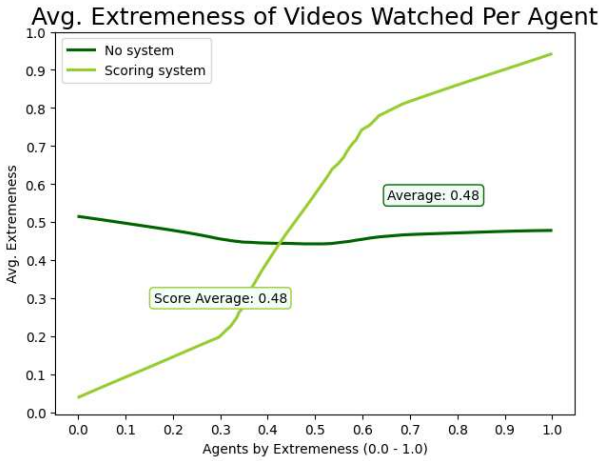


Fig. 10: Extremeness watched with no system and scoring system.

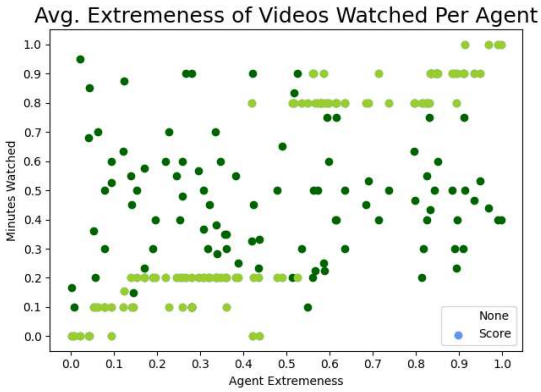


Fig. 11: Scatter plot version of Fig. 10, to highlight the bifurcation.

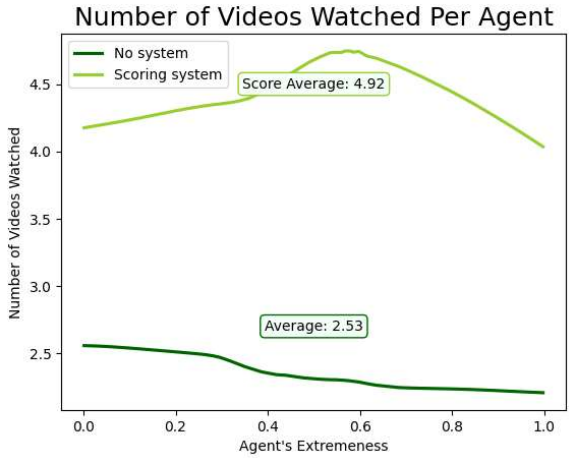


Fig. 12: Avg. number of videos watched with no system and scoring system.

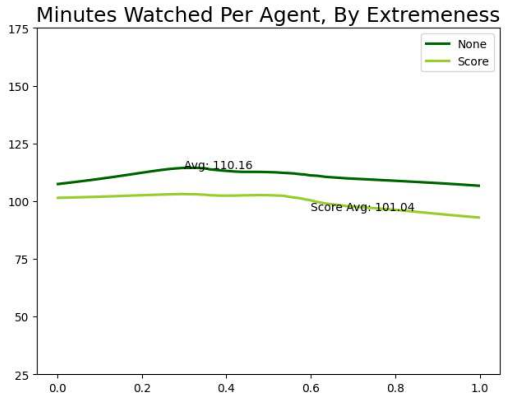


Fig. 13: Avg. minutes watched with no system and scoring system.

VII. Discussion and Wrap-Up

All of these graphs and all of this code prove my initial hypothesis---that scoring systems like YouTube's push users to watch more politically extreme content than they would in the absence of any recommendation system. In Fig. 10, users without the influence of the video scoring system watch videos of about average extremeness (0.56, compared to 0.5 being "average," non-partisan videos. However, with the scoring system, users already close to the political extremes (0.0 and 1.0) are pushed even further towards their personal extremes, and watch less mainstream videos—or videos that differ at all in extremeness from their tastes.

Because the scoring system is rewarding videos with an extremeness ≤ 0.2 or ≥ 0.8, we can see that most users are watching videos around those extremeness, *regardless of their own personal preferences!* Obviously, part of that is because it's coded to do that to some extent, but it was really interesting to see how even the more middle-road users got pushed to extreme content (which is the whole point of this project).

In Fig. 12, users with the scoring system watch almost double the amount of videos than without. As the scoring system rewards shorter videos, users can fit more videos into a watching session—allowing social media companies to show users a larger variety of inflammatory content to engage with. This further aligns with the rise in "short form" content on social media websites, such as TikTok, YouTube's Shorts, and Instagram's Reels.

So, what next? I feel that this project could go two ways. One is to try and model this phenomenon with real video and user data from YouTube. Because I am a one-person research team, I can only do so much, so this is all simulated data. Existing research has attempted to study this phenomenon, but opinions seem to be split on whether or not YouTube is doing any harm. The other path could be really refining this simulation to get it as close as possible to how we know YouTube operates. This could mean switching to a deep learning model, adding more attributes to videos and users, or many other coded add-ons. That way, if these findings are repeated, it would be much easier to claim that YouTube's systems are pushing users towards politically extreme content, and ideally make someone do *something* about it. Even just highlighting these issues is, I think, an important step.

VIII. Why Care?

Nobody has asked me this yet, but I feel it's important to answer. In my line of work thus far, I've spent hours upon hours studying the real impacts of engineering systems, specifically computational ones. Data and computing are not as objective as we think, and somebody has to be there to make sure digital systems aren't causing offline harm to real people. Through work like this, and as cheesy at it sounds, I hope I can begin to be that somebody.

Acknowledgments

Chiefly, I'd like to thank my cats, Josie and Charlotte Rabeno, for their invaluable contributions to the research process. As for human contributors, I wish to thank Dr. Larry Snyder (probably the only one reading this far), my research advisor for the last two years. My academic advisor in the IDEAS program, Bill Best, who has instilled in me the importance of asking about ethical obligations. My family and friends, for helping me through my undergraduate journey. And you, whoever is reading this!

References

[1] S. Atske, "2. A closer look at the channels producing news on YouTube – and the videos themselves," *Pew Research Center*, Sep. 28, 2020. Accessed: Apr. 22, 2024. [Online]. Available: https://www.pewresearch.org/journalism/2020/09/28/a-closer-look-at-the-channels-producing-news-on-youtube-and-the-videos-themselves/

[2] Defunctland, "Disney's FastPass: A Complicated History," *YouTube*. Nov. 21, 2021. Accessed: Apr. 22, 2024. [Video]. Available: https://youtu.be/9yjZpBq1XBE?si=hl7O1xF9YDkuGoDs

[3] S. Hawkins, D. Yudkin, M. Juan-Torres, and T. Dixon, "Hidden Tribes: A Study of America's Polarized Landscape," 2018. Accessed: Apr. 22, 2024. [Online]. Available: https://hiddentribes.us/media/qfpekz4g/hidden_tribes_report.pdf

[4] P. Covington, J. Adams, and E. Sargin, "Deep Neural Networks for YouTube Recommendations," in *Proceedings of the 10th ACM Conference on Recommender Systems*, Sep. 2016. Accessed: Apr. 29, 2024. [Online]. Available: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45530.pdf

[5] S. Rabeno and L. Snyder, "tools-of-justice repository page," *GitHub*, Apr. 28, 2024. https://github.com/snyder-research-group/tools-of-justice (accessed Apr. 28, 2024).