Why Do Content-Recommending Social Media Algorithms Cause Radicalization?

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

James Joseph Connors III

Spring 2023

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

MC Forelle, Department of Engineering and Society

Introduction

Social media may be one of the most influential pieces of technology in the modern age. Since the inception of the internet, an enormous amount of social discourse, commercial business, and information sharing has shifted to take place on these websites and applications. About seven in ten Americans use social media to connect with one another (Pew Research Center 2021), and about half of all Americans get news on social media at least sometimes (Walker 2021). Worldwide, 5.03 billion people use the internet, with 4.7 billion of those people also using social media (Statista). With this level of global outreach, any decisions made in the design of these applications can affect the lives of billions of people. For this reason, it's vital that their design and impacts are closely scrutinized.

Of course, despite this ubiquity and importance, the effects that social media has on its users and societies worldwide are highly controversial. One particularly worrying aspect of social media is the dubious impact made by the algorithms that these companies use to determine what a user sees when they engage with a social media platform. To give a broad summary, companies like Facebook or Twitter derive their profit from keeping users engaged with their platform, as the longer a user stays engaged, the more exposure advertisements receive, which generates them revenue. In order to keep those users engaged, these companies collect data on what their users see and interact with and feed it into algorithms which predict what content would keep them on the website in the future. Then, the user's feed is populated with that content (Kim 2017). Ultimately, these algorithms define what's shown to users of social media, which means that they can have a tremendous social impact. As the arbiters of what content is shown on these platforms, content-recommending social media algorithms can greatly affect the news, conversations, and views that billions of people see every day.

Unfortunately, research into these algorithms has revealed a number of problems. Many critics of content-recommending algorithms have pointed out that they have an alarming tendency to recommend content that is inflammatory and can lead to political and ideological extremism (Deibert 2019; Kim 2017; Ribeiro 2020). Others discuss the negative effects that these algorithms have on news media (Peterson-Salahuddin 2020) or how they can be used by authoritarian governments as extremely effective political tools (Bradshaw 2019), among many other issues. Ultimately, there is a copious amount of research that points to the fact that content-recommending social media algorithms cause radicalization and polarization in many of their users. However, less has been said on why these algorithms are designed in this way. In order to fix some of the problems with these algorithms, we must first understand the reasons that these algorithms are designed in the way that they are, and why attempts to quell these problems have so far been unsuccessful. Following that, we might get a glimpse into how to better improve them in the future.

In this thesis, I posit that social media algorithms are designed in such a way as to cause radicalization and polarization because the psychological forces that lead to these problems are closely aligned with the goal of maximizing profit for social media websites. Further, I posit that attempts currently in place to curb the problems with these algorithms are generally insufficient due to a core conflict with the design of these algorithms and the websites they support. Through this analysis, I aim to gain a greater understanding of these algorithms and develop a few recommendations for future research that are more promising for curbing their problems.

This thesis will begin with a literature review discussing past research on the controversial impacts of these algorithms on society, with a special focus on their tendency to cause radicalization and polarization. Following that, I'll discuss my methodology, which will be

to conduct a discourse analysis on what researchers are saying on the reasons why these algorithms are designed in this way, as well as why past attempts to fix these problems have failed. This discourse will also review how past attempts to solve the problems with these algorithms have failed, and will highlight where future research might improve on them. Through this analysis I will find why content-recommending algorithms are designed in a way that causes radicalization, why past attempts to fix these problems have failed, and where future steps might be taken to redesign them for the better.

Literature Review

Social media algorithms have long been controversial for their tendency to foster extremism and polarization. Deibert (2019) coined three painful truths about these algorithms, which are that they're built around data surveillance and spying, that they're designed to be addictive, and that their attention-grabbing tendencies propel authoritarian practices that aim to "sow confusion, ignorance, prejudice, and chaos." (Deibert 2019 p.8). On the topic of authoritarianism, Bradshaw (2019) found that social media algorithms were manipulated by the world's governments to shape public attitudes, and in some cases were used as a tool to suppress human rights, discredit political opponents, and drown out dissenting viewpoints. News has also been cited as suffering under the algorithm. After surveying journalists and editors, Peterson-Salahuddin (2020) found that journalists often realized that when writing, they needed to negotiate between practices that would benefit their chances at having their article be recommended by these algorithms and practices that would fit their concepts of newsworthiness or journalistic autonomy. Perhaps most famously, in extreme cases these algorithms can even lead to radicalization of individuals to dangerous or fringe viewpoints. Ribeiro (2020) aimed to see if this narrative of radicalization pathways held weight on YouTube, and found this narrative

to be largely true. What's worse, van Eerlen (2017) found that once radicalization has taken hold, that it's profoundly difficult to reverse, with the far better strategy being to simply prevent it in the first place.

Of course, that's not to say that the discussion on these algorithms is always a monolith. Johnson (2019) found that extreme views can come into a society even when everyone has the same information and resources, and found that certain social media algorithms designed to reduce division can actually worsen it. Then there's the discussion of the filter bubble, an idea first popularized by Pariser (2012), which stated that the personalization of social media and search algorithms would leave users in their own "bubbles" where they would never be challenged by dissenting viewpoints and instead only see their own views parroted back at them. This idea is popular and has been built on by some authors proposing ways to visualize this bubble (Nagulendra 2014) or positing solutions to the problem (Bozdag 2015), but the idea has also been challenged by other authors who cite a lack of evidence in the claim (Bruns 2019, Dahlgren 2021, Haim 2017). Törnberg (2021) also discusses the more recent idea that polarization isn't caused by bubbles or echo chambers, but rather by the emergence of partisanship as a social identity. Ultimately, the current politics of social media algorithms are highly complex and varied, but it's safe to say that they do encourage a society in which people are driven to ideological and political extremes.

On the topic of a society driven to extremes by social media, it's worth mentioning the STS framework that inspired the creation of this thesis. In his 1980 essay "Do Artifacts Have Politics?", author Langdon Winner discusses the idea that technical objects have inherent political properties, and both embody and encourage certain political ideologies or ideas. Deep into the essay, Winner proposes the idea that some technologies promote certain sociological

systems because those technologies are either only able to function in certain sociological systems, or at least benefit immensely from those systems. Winner gives the example of nuclear power plants, citing that the technology of nuclear power all but requires an authoritarian governmental body of some kind to be put in place in order to prevent nuclear disasters (Winner 1980). With just how pervasive problems of polarization and radicalization are across social media algorithms, it makes one wonder if social media is one of these kinds of technologies. In theory, if radicalized users make social media companies more money, and these algorithms in turn attempt to maximize profits, it could be argued that these algorithms both produce a more radicalized society and benefit from a more radicalized society. In turn, these algorithms create a sort of feedback loop where they become increasingly profitable and increasingly problematic for society as a whole. This line of reasoning was what inspired this thesis's look into the reasons why these algorithms are designed in the way that they are, as well as what methods might help to change them for the better. Using Winner's work as a framework to guide the discussion, this thesis will analyze whether the evidence shows that these algorithms require a divided society in order to function, or whether there is some hope that they could exist without both promoting and benefitting from high levels of division and radicalization.

Methods

Ideally, research for this thesis would rely on primary sources, but due to the nature of these algorithms and these companies, such an analysis isn't entirely feasible. Any analysis of the algorithm directly is off the table, as content-recommending algorithms on these platforms are almost universally black-box machine learning algorithms. In this context, fully understanding exactly why an algorithm picked a specific piece of content to recommend is all but impossible, even for the developers of these algorithms. Certainly, the designers of the algorithms likely had

design goals for what trends the algorithms would follow, but getting a direct quote or interview with the designers of these algorithms is also unlikely. Unfortunately, most of these designers are under non-disclosure agreements or are otherwise incentivized to keep quiet.

Because of these restrictions, this analysis will focus primarily on secondary sources, in an attempt to perform a discourse analysis on literature surrounding these algorithms. To begin, I will examine sources pertaining to the reasons why social media algorithms are designed in a way that produces extremism. These will include studies on the psychology of radicalization in general as well studies examining radicalization pipelines in social media and how those psychological tricks pertain to social media algorithms. Following that, I will examine literature on the various methods that have been put in place by these companies and the public in order to combat social media radicalization and polarization. Finally, I will review sources that suggest or show where future improvements can be made on these algorithms, if any exist. Through these three points of analysis, I hope to gain a fuller understanding of the decisions made around these algorithms and the impacts of those decisions. This way, I will determine why it is that content-recommending algorithms are designed in a way that promotes extremism, why current methods to alleviate these problems have struggled to do so, and where we can go from here.

Results/Analysis

Overall, the literature supports the idea that social media algorithms are designed to cause radicalization because the psychological forces that drive engagement are also primed to sow seeds of radicalization and polarization. As mentioned earlier, the stated goal of many social media algorithms is to drive engagement, which essentially means that social media companies want their users to be actively using their websites for as long as possible. The prevailing idea is that the longer a user is paying attention to and engaging with a social media platform, the more

they'll watch and click on advertisements that make the company money (Kim 2017). Thus, social media companies tune their algorithms to encourage long, active, and attention-grabbing sessions on social media websites (Kim 2017). Unfortunately, radicalizing and polarizing content is extremely good at driving engagement, which means that content of that ilk is likely to be pushed by these algorithms. For example, Rathje (2021) found that posts on Facebook and Twitter that provoked out-group animosity (meaning content shared in one political group that disparaged their rivals) was shared or retweeted about twice as often as posts about the in-group. These reactions are engagement, and as such show one way that psychological forces that drive humans to engage with social media (such as tribalism) also tend to drive polarization.

Other social media websites also share these problems largely due to their focus on engagement. Ribeiro (2020) viewed a common narrative in the statements of non-profits and the media, which was the idea that YouTube created radicalization pathways that would systemically cause users to progress towards more extreme content on the platform. Through an examination of over 330,925 videos posted on 349 channels, Ribeiro found this common narrative to be largely true. Through analyzing comments on videos with varying levels of radicalization, Ribeiro found that users consistently migrated from less extreme to more extreme content over time. Similar work was done by Boucher (2022), who performed a smaller experiment using four TikTok profiles that interacted with conservative content, all of which were themselves shown increasingly more extreme videos recommended to them over time. These "radicalization pipelines", where social media algorithms will consistently show their users more and more radical content over time, are especially worrying, as they mimic the psychological tricks that radicalizing groups have used throughout human history. Boucher (2022) cites how these online pipelines and the communities that form around them work to slowly normalize or banalize

fringe and radical ideas for their members, all while providing a sense of community and acceptance for the people inside those communities. This is similar to how radicalization has occurred in the real world in the past, but social media has allowed it to happen at a much more widespread scale and a much faster rate. At their core, these tactics of community building, identity formation, and a slow introduction of radical ideas are very compatible with the goals of these social media websites, which is to increase engagement. This causes these websites to promote these spaces, causing radicalization. Overall, the reason that social media algorithms are designed in a way that promotes extremism is that the psychological forces that lead to these problems are closely aligned with the goal of maximizing profit for social media websites.

While social media websites' goals make them very compatible with radicalizing tendencies, it would theoretically be possible for them to alleviate this problem by purposefully policing or otherwise regulating harmful or radicalizing content, even if their underlying algorithms are trying to promote it. Unfortunately, attempts currently in place to curb the problems with these algorithms are generally insufficient, and the reasons for this stem from deep-seated inherent problems with the technology as a whole. Most of these websites have some sort of policy that restricts what kinds of harmful content can be posted on their website. Many of these policies are well-intentioned, but they can be borderline impossible to fully enforce. For example, Becker (2021) explores the 'enforcement gap' at TikTok, which is the huge number of videos that are blatantly against TikTok's terms of service but remain up on the platform for extensive periods of time. This included videos from violent extremists, neo-Nazis, and other white supremacist groups, which in some cases expressed explicit support for known extremists, terrorists, and mass shooters. Even the most radical content is difficult to police because the design of these websites creates inherent problems for such policework. Too many

people post content to these websites for any reasonable numbers of humans to manually review and approve content, and cases like TikTok show that even with other algorithms and tools to identify potentially problematic videos, many can slip through the cracks.

Social media websites can and do employ a variety of moderation methods, which can include both human-powered methods like moderators and more algorithmic-focused methods. Sheng (2022) provides an overview of these methods, as well as their strengths and weaknesses. As mentioned before, manual moderation techniques are often limited in their scope, as the sheer scale of social media can involve far too much content for workers paid by these companies to sift through. Content moderation algorithms, then, are used to handle this scope problem. These moderation algorithms, however, are also flawed. Sheng (2022) cites how they can be biased against certain groups depending on their training, how they can often struggle to grasp a complete contextual understanding, and how they are completely non-transparent. These moderation algorithms are typically made as black-box machine learning algorithms, which means that any decision they make can't be fully understood, often leaving users at a loss for why their posts were taken down or why harmful content remains on a website. It's also notable that the moderation policies themselves, even if they were perfectly enforced, can't always solve the problems with social media algorithms favoring radical content. After all, radicalizing pathways as described by Ribeiro (2020) and Boucher (2022) contain a continuum of increasingly radical content, with some earlier steps of that pathway being relatively mainstream, non-radical ideas. Where then should a social media company draw the line with what ideas are too radical to be allowed? There isn't a clear answer. What's worse, these companies have to be careful about what and how much they remove, as Coscia (2022) finds that policing content online can sometimes even lead to backlash and more polarization for both users and news

sources. Finally, even relying on occasional intervention by users is shoddy at best; Kim (2020) found that comments correcting misinformation online were judged as reasonable primarily based on the tone of the comments, rather than by their content or the actual truth. Overall, due to the nature of the technology, it is extremely difficult to police harmful content that's proliferated by social media algorithms, which further exacerbates the problems with the technology.

This provides a worrying outlook for the technology of social media, but the literature doesn't necessarily suggest that the technology is completely unsalvageable. Perhaps most intriguing is the work of Yarchi (2020), which suggests that political polarization on social media is not a unified phenomenon, but instead differs from application to application. When studying Facebook, Twitter, and WhatsApp on three different aspects of polarization, the researchers found that different platforms were very different in which aspects of polarization they exhibited, as well as the severity of those aspects. This points to a flaw in the idea of social media radicalization, which is that all of these websites are admittedly different, with differences in their algorithms, content, and surrounding design. It's possible that further comparisons between multiple social media websites and the algorithms that run them could lead to a greater understanding of how different platforms deal with radicalization, and how they could all improve. Further, while the design of these algorithms is of course meant to produce as much revenue as possible for their companies, it's worth noting that social media companies are currently making billions in revenue each year (Ku 2022), so it's reasonable that with some regulation, they could be forced to make algorithms that are less profitable but also less harmful to society and still function as a successful business, just with smaller profits. Imana (2022) proposes an alternative method for the regulation and auditing of social media algorithms for the

public's benefit. In this method, third-party auditors would be given limited, privileged access to relevance estimators that they could use to better audit and understand the decisions made by these algorithms. The study also introduces ideas on how to protect against risks of privacy loss and the discovery of proprietary business secrets, which are issues that have held up such audits in the past. Schemes similar to this proposal would be a good start for regulatory action on these algorithms. Ultimately, the literature does not support the idea that all hope is lost on improving these algorithms, even if any attempt at improvement may be difficult.

Conclusion

As mentioned before, Langdon Winner's "Do Artifacts Have Politics?" posits the idea that some technologies require a certain political structure in order to function. At its core, this thesis aims to find whether the technology of social media is so compatible with a divided and radicalized society that it requires such a society to exist in order to fulfill its goals. Certainly, the evidence shows that social media algorithms heavily benefit from division and radicalization when attempting to make money, and that the technology in turn has a tendency to promote such qualities in its users. The evidence also shows that the technology doesn't play well with attempts to dissociate it from these ties to radicalization, with many current solutions acting more as band-aids than true fixes.

Despite all of this troubling information, however, this thesis still finds that there is not yet sufficient evidence to say that these algorithms simply must require or promote a divided and radical society. The technology is not necessarily unsalvageable; future research on the differences between these algorithms and platforms could provide notable insights into possible improvements, and there is reason to believe both that regulation could be possible, and that regulation could be financially feasible. It's vital that future research further examine these areas

in order to improve the political impacts of these algorithms. As of right now, it's too early to call whether social media algorithms can be fixed. There are certainly huge problems in their current design, and a great deal of research must still be done in order to improve them. Still, there is some hope that in the future, these algorithms could exist without promoting radicalization, provided that we learn to regulate, change, and tune them correctly.

Works Cited

Becker, A. (2021, August 24). *On TikTok, misogyny and white supremacy slip through 'enforcement gap'*. The 19th. Retrieved March 15, 2023, from https://19thnews.org/2021/08/tiktok-misogyny-and-white-supremacy-slip-through-enforcement-gap/

Boucher, V. (2022). *Down the TikTok Rabbit Hole: Testing the TikTok Algorithm's Contribution to Right Wing Extremist Radicalization* [Graduate Thesis, Queen's University]. http://hdl.handle.net/1974/30197

Bozdag, E., & van den Hoven, J. (2015). Breaking the filter bubble: Democracy and design. *Ethics and Information Technology*, *17*(4), 249–265. https://doi.org/10.1007/s10676-015-9380-y

Bradshaw, S., & Howard, P. (2019). The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation. Project on Computational Propaganda. https://demtech.oii.ox.ac.uk/research/posts/the-global-disinformation-order-2019-global-inventory-of-organised-social-media-manipulation/

Bruns, A. (2019). Filter bubble. *Internet Policy Review*, *8*(4). https://doi.org/10.14763/2019.4.1426

Coscia, M., & Rossi, L. (2022). How minimizing conflicts could lead to polarization on social media: An agent-based model investigation. *PLOS ONE*, *17*(1). https://doi.org/10.1371/journal.pone.0263184

Dahlgren, P. M. (2021). A critical review of filter bubbles and a comparison with selective exposure. *Nordicom Review*, *42*(1), 15–33. https://doi.org/10.2478/nor-2021-0002

Deibert, R. J. (2019). Three painful truths about social media. *Journal of Democracy*, *30*(1), 25–39. https://doi.org/10.1353/jod.2019.0002

Haim, M., Graefe, A., & Brosius, H.-B. (2017). Burst of the filter bubble? *Digital Journalism, 6*(3), 330–343. https://doi.org/10.1080/21670811.2017.1338145

Imana, B., Korolova, A., & Heidemann, J. (2022). Having your Privacy Cake and Eating it Too: Platform-supported Auditing of Social Media Algorithms for Public Interest. *ArXiv*. https://doi.org/https://doi.org/10.48550/arXiv.2207.08773

Johnson, N. F., Manrique, P., Zheng, M., Cao, Z., Botero, J., Huang, S., Aden, N., Song, C., Leady, J., Velasquez, N., & Restrepo, E. M. (2019). Emergent dynamics of extremes in a population driven by common information sources and new social media algorithms. *Scientific Reports*, 9(1). https://doi.org/10.1038/s41598-019-48412-w

Kim, J. W., & Masullo Chen, G. (2020). Exploring the influence of comment tone and content in response to misinformation in social media news. *Journalism Practice, 15*(4), 456–470. https://doi.org/10.1080/17512786.2020.1739550

Kim, S. (2017, December). Social Media Algorithms: Why You See What You See. Georgetown *Law Technology Review*, 2, 147–154.

Ku, D. (2022, October 6). *Which social media platforms make the most revenue?* PostBeyond. Retrieved March 15, 2023, from https://www.postbeyond.com/blog/revenue-per-social-media-user/

Nagulendra, S., & Vassileva, J. (2014). Understanding and controlling the filter bubble through Interactive Visualization. *Proceedings of the 25th ACM Conference on Hypertext and Social Media*. https://doi.org/10.1145/2631775.2631811

Pariser, E. (2012). *The filter bubble: What the internet is hiding from you*. Penguin Books.

Peterson-Salahuddin, C., & Diakopoulos, N. (2020, July 10). Negotiated Autonomy: The Role of Social Media Algorithms in Editorial Decision Making. *Media and Communication*, 8(3), 27–38. https://doi.org/10.17645/mac.v8i3.3001

Pew Research Center. (2021, April 7). *Social Media Fact sheet*. Pew Research Center: Internet, Science & Tech. Retrieved October 27, 2022, from https://www.pewresearch.org/internet/fact-sheet/social-media/

Rathje, S., Van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences, 118*(26). https://doi.org/10.1073/pnas.2024292118

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2020, January 22). Auditing radicalization pathways on YouTube. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3351095.3372879

Sheng, C. (2022). Automated Content Moderation. *Georgetown Law Technology Review*.

Statista Research Department. (2022, September 20). *Internet and social media users in the world 2022*. Statista. Retrieved October 27, 2022, from https://www.statista.com/statistics/617136/digital-population-worldwide/

Törnberg, P., Andersson, C., Lindgren, K., & Banisch, S. (2021). Modeling the emergence of affective polarization in the Social Media Society. *PLOS ONE*, *16*(10). https://doi.org/10.1371/journal.pone.0258259

van Eerten, J., Doosje, B., Konjin, E. A., & de Graff, B. (2017). Developing a social media response to radicalization: The role of counter-narratives in prevention of radicalization and de-radicalization. *ResearchGate*.

Walker, M., & Matsa, K. E. (2021, September 20). *News consumption across social media in 2021*. Pew Research Center's Journalism Project. Retrieved October 27, 2022, from https://www.pewresearch.org/journalism/2021/09/20/news-consumption-across-social-media-in-2021/

Winner, L. (1980). Do Artifacts Have Politics? *The MIT Press*, 121–136.

Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2020). Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication*, *38*(1-2), 98–139. https://doi.org/10.1080/10584609.2020.1785067