For You Page or For Corporate Profits: How Recommender Systems are Harmfully Designed

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

Pawan Jayakumar

Spring 2024

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

Caitlin D. Wylie, Department of Engineering and Society

"We've been fueling this fire for a long time and we shouldn't be surprised it's now out of control" (Suderman, 2023). These are the words of a Facebook employee in response to the capitol riots that took place on January 6th 2021. Not long after, a whistle blower disclosed thousands of internal documents to the Securities and Exchange Committee. The papers revealed

how Facebook willingly allowed the spread of misinformation and was incapable of moderating extremist groups on its platform. Social media companies encourage people to upload content and spend time interacting with their platforms in order to gain revenue from advertisers. Recommender systems are a major feature found in all of the biggest social media platforms such as the Twitter (now known as X) timeline or TikTok's "for you page" which filter and display content it thinks is relevant for the user. Recommender systems cause significant harm to their users because they are designed to be addicting, provide biased recommendations, and raise data privacy concerns. The rest of this paper will first provide a brief overview of frameworks and research methods used, and then will go over each of these issues in detail.

This paper will be utilizing the Actor Network Theory (ANT) , which was first introduced by John Law (Law, 1992). ANT is an STS theoretical framework which posits that all of society exists in a network of human and non-human actors and that any social behavior can be explained through a careful examination of the interactions between these actors. This framework is valuable for its emphasis on the role which non-human actors (ie. technologies such as recommender systems) play on social dynamics. Using ANT, this paper will analyze documents and case studies.

In this problem, the key actors are social media companies, advertising agencies, social media users, and the recommendation system itself. Social media companies exist to generate revenue by providing other companies a platform on which they can advertise on. By promising these companies that many people will see their ads, they can charge quite a lot of money for it. According to Statista, social media advertising was over a 207 billion dollar market and is expected to show an annual growth rate of 3.86% for the next four years (Statista, 2023). This type of relationship between advertisers and social media companies encourages the need to keep users on the app or platform for as long as possible. The more videos a user watches on TikTok, the more advertisements they will be exposed to, and what better way to increase watch time than to entice users and make them addicted.

Recommender systems and addiction

Social media companies design their platforms using the Hook model, created by Nir Eyal, to create and enforce addictive habits into its users in order to increase usage and retention (Eyal, 2023). Nir Eyal is a Stanford MBA graduate who has worked in the advertising industry for years and has founded and sold two companies. The Hook model is a four step process that is implemented by applications to create desirable habits in its users. First there is some kind of trigger (such as a notification or simply seeing the app icon) which the user learns to associate with wanting to open the application. Then there is the action which the user takes in response to the trigger. In the case of TikTok, this is watching videos and swiping to see a new one. After taking the action, the user is presented with a schedule of variable rewards. Not every video a user watches on TikTok is guaranteed to elicit a positive response. Finally, the user is asked to make some kind of investment which helps the company as a sort of payment for the variable rewards received (Eyal, 2023). For example, every few videos on TikTok is actually an advertisement. Twitter employs this model in a similar fashion, just replacing videos with tweets. Variable schedules of reward are one of the most powerful tools that companies use to hook users. Research shows that levels of dopamine surge when the brain anticipates receiving a reward (Eyal, 2023). Recommendation systems don't always show you content that you particularly like, but when it does, the surge of dopamine is very powerful. Variable rewards is the same reason why gambling addiction is such a prevalent issue. While the machine is designed for the house to win in the long run, the small fixed chance of hitting the jackpot tricks

our brain into believing that the next one could be it. Despite knowing that it is objectively not worth pulling the lever, variable rewards can incentivize people to do so anyway because they are willing to pay for the excitement. Even the user-interface design of these recommendation systems enables addictive behavior. Immediately upon opening the app, users are presented with a video from their "For You" page. This means the user does not have to spend time thinking about what to watch. Tiktok also makes it very easy for users to continue watching videos, as a simple upwards swipe seamlessly loads and starts a new video from the recommender. Through variable rewards and a carefully designed user interface, it should come to no surprise that social media platforms like TikTok and Twitter have over 1.5 billion and 600 million monthly active users respectively (Dixon, 2024). Such a large user base can have profound effects on society.

Recommender systems also contribute towards addiction by creating social events. The virality of certain tweets or Tiktok videos causes a wave of similar types of content to be uploaded in hopes of replicating the success. These include popular jokes or memes, dances, outfits, recipes, life skills, and more. Since so many people are exposed to these trends, they talk about them in daily conversations. Some users may feel that by not regularly checking Tiktok or Twitter, they cannot stay up to date with the latest trends. In a survey conducted among over 2000 people ranging from eight to nineteen years old, 56% of respondents in a survey said they reported this fear of missing out (Dixon, 2013). The ease of finding content, the exciting rewards behind it, and the prospect of going viral by posting videos mean there is always a convenient access to dopamine. According to the National Institute of Health, one of the most reputed biomedical research groups in the world, dopamine has been linked to addiction (National Institute of Health, 2022). While it is vital for bodily function, too much dopamine has been linked to addiction because neurotransmitters adapt to a steady influx of dopamine by becoming less sensitive. This results in other activities feeling less rewarding and we lose motivation to do them (National Institute of Health, 2022).

Mental health concerns related to recommendation systems go far beyond addiction. Amnesty International is a humanitarian organization which has investigated the harm these personalized systems create. In collaboration with the Algorithmic Transparency Institute, they ran experiments with Tiktok's recommendation system by modeling fake users to scroll through videos on the platform. Within hours of this process, they found that "TikTok's 'For You' feed can easily draw children and young people who signal an interest in mental health into 'rabbit holes' of potentially harmful content, including videos that romanticize and encourage depressive thinking, self-harm and suicide (Amnesty International, 2024). TikTok risks exacerbating children and young people's struggles with depression, anxiety and self-harm, putting young people's mental and physical health at risk." All it takes is one interaction with a topic and the recommender system begins to present more and more, leading to a downward spiral. In addition to explicitly depressing content, recommender systems can promote content which exacerbates insecurities. A study conducted in 2023 analyzing over 200 teens in London found that body dysmorphia was significantly and positively correlated with social media use (Gupta et al., 2023). Many viral trends either start or become popular due to conventionally attractive individuals, and leads to comparisons being drawn between users and these individuals whose job it is to look as good as possible. Unrealistic standards being presented as normal can have damaging effects on viewers. While not directly causing body dysmorphia and other insecurities, recommender systems often enable the process by bombarding users with content which can cause it just because the user interacted with one such piece of content. The quickness

of these systems to trap users in a bubble of certain types of content is the source of the next big issue with recommender systems.

Recommender systems and bias

Recommender systems provide biased recommendations by taking shortcuts and making assumptions about user preference. At the core of all of these systems is some sort of user preference modeling algorithm. Even without explicitly mentioning what we enjoy, these algorithms learn our preferences through our actions on the website. For example, clicking on someone's tweet and/or commenting on it (generally referred to as interacting with it) will inform the algorithm of the type of content a user prefers and will have a higher chance of promoting tweets from that author (Twitter, 2023). This model employs many assumptions which heavily affect the quality of the recommendation. In the earlier example, it can be the case that the user viewed their tweet but didn't like it and commented something negative. It is also possible the user only really liked the one tweet this author made but wouldn't generally want to see their future tweets. Either way, it highlights the potential for disconnect between what the user wants and what the recommender system thinks they want.

This disconnect is especially an issue for new users of a platform because the recommender system has no idea what the user prefers. When humans are found in similar situations, they often employ stereotypes as a first best guess. Recommender systems aren't much different: the Twitter timeline and many other systems employ a method called collaborative filtering (Twitter, 2023). Take two users Alice and Bob, and two sweets: cookies and donuts. If Alice and Bob both purchase cookies, the recommender system sees there is some correlation between their interests. If Alice then purchases donuts, the system would predict that Bob may also purchase donuts and recommend it to him. This is a very simple example of collaborative filtering, but one can imagine it being scaled up to specific groups of users (perhaps all fans of a particular sports team) and specific groups of products (sports merchandise). Collaborative filtering is what leads to stereotyping. A team from Netflix's research and development division have studied this issue in great detail. "If preferences for a set of items are anti-correlated in the general user population, then those items may not be recommended together to a user, regardless of that user's preferences and rating history" (Guo et al., 2021). Netflix is one of the largest streaming service providers worldwide, offering thousands of movies and TV shows to their subscribers. They are incentivized to build a high quality recommender system for them because it directly contributes to their success: if subscribers can't find the content they like, they will quit the service. While affiliated by Netflix, these researchers are still credible, as they have collaborated with professors in academia, and their work has been cited over 20 times. By placing users into arbitrary bins, recommender systems are not able to capture the uniqueness of each user. Just because most sports fans only purchase their favorite teams merchandise doesn't mean someone can't prefer having multiple team jerseys. It can also mean recommender systems can perpetuate harmful stereotypes about certain social groups.

We're not out of the woods even if the recommender system has accurately captured part of the user's preference. In fact, I'd argue that this is the case where they present the highest potential for abuse. The constant cycle of users interacting with items and recommender systems showing similar items creates a positive feedback loop. Two researchers from Australia National University and the Alan Turing Institute have shown that recommender systems don't consider that by giving recommendations, they are actively influencing the user's preference (Evans, 2022). Nature is devoid of positive feedback loops for a reason: they tend to spiral out of control. At the end of this spiral lies echo chambers, conspiracy theorists, mis-information and political

extremism. The scary thing is this slippery slope can take effect over time and feel natural to the user. Separate institutions have come to the same conclusion: social media users are more likely to hear only the viewpoints that align with their own, an effect often referred to as an "echo chamber." (VCU, 2023) These systems have recently gone under heavy criticism for promoting the creation of filter bubbles, lowering the diversity of information users are exposed to and the social contacts they create (Fernandez, 2020). The dangers of political extremism are not hard to imagine, because their effects have led to some of the largest events in modern history. The January 6 riots following Biden's election into office showcases how social media can be used to coordinate extremist ideology (Anti Defamation League, 2022). When an individual is part of an echo chamber, they are constantly bombarded with views that reinforce their own. This easily allows misinformation to spread and become accepted as the truth. Even if posts containing misinformation are taken down, it's usually too little too late. Recommender systems enabling the spread of misinformation allows individuals to find and/or recruit others with similar extremism views. During the covid-19 pandemic, partly due to distrust in government and partly due to the speed at which they were developed, vaccines became a controversial topic. The conspiracies ranged from the vaccine having harmful effects all the way to implanting microchips to track citizens (Anti Defamation League, 2022). While a healthy dose of skepticism is important, it has become abundantly clear that these vaccines do not come with microchips, nor do they cause autism and the fact that people continue to believe they do shows how potent echo chamber ideologies can be.

Looking back at this issue from the perspective of ANT, there is a careful balancing act that must be done by social media companies. On one hand, they do want to prevent extremism from occurring. Public perception about platforms that support extremism will seldom be positive. On the other, they don't want to lose users by forcing them to view content that they don't care about, just because it promotes diversity of information. Twitter has taken measures to try and combat this by promoting diversity in their timeline algorithm as well as empowering the community to add context to tweets that present misinformation (Twitter, 2023). While promising, this feature does not necessarily reduce the amount of misinformation people see. Researchers from the University of Luxembourg conducted a large scale study of Twitter community notes and found that it failed to prevent interactions with the misinformation in the early stages where virality is most likely to occur (Chuai et al., 2023). Unfortunately, people view and share misinformation before the community notes correct the tweet. It is hard to manage such a large scale of misinformation being produced and circulated. Until technology improves to the point where information can be fact checked in real time, it will be hard to avoid this issue. Even then, it is unclear whether a real time fact checking system targets the issue at its root. Society in general must be taught to employ critical evaluation of the sources they obtain their information from. Misinformation is at an all time high and the emergence of generative machine learning models which can be used to imitate speech or create fake videos of target individuals only contributes to that issue. It is up to individuals to put in the work to be cautious about what they see.

Recommender systems and privacy

There is a tradeoff between privacy and effectiveness because the more personally identifiable information the recommender system has, the better recommendations it can make (Friedman, 2015). As an example, if someone routinely purchases a certain product every month,

the system can be pretty confident that they will do so again and will remind the user to purchase that product. This immediately raises privacy concerns: how are these platforms collecting, storing and using personal data? All three of these actions come with risks that have not been adequately addressed. Social media companies have a poor record of being transparent about how and what data they collect on users. While technically found in the terms of service agreements for using the platform, these documents are so long that most users don't bother reading them. According to a survey conducted in 2019 by the Pew research center, only 20% of US adults said they always or often read the terms and conditions. Even among those who do, only 22% read it all the way through and only 13% say they "understand a great deal" about them (Pew Research Center, 2019). This is quite concerning given the plethora of things social media companies collect on users. According to TikTok's privacy policy, they may collect biometric identifiers and biometric information as defined under U.S. laws, such as faceprints and voiceprints, from content uploaded (TikTok, 2024). This type of data is about as personally identifiable as it gets. They also gather approximate location data by automatically collecting your IP-address, SIM card data, and analyzing content you upload (TikTok, 2024). This really is not info you want in the hands of hackers, but that might just be what happens.

While the act of storing data itself may not present a privacy concern, that is assuming only certain parties have access to it. Unfortunately, data breaches are quite common. Infamously, Facebook was under fire for allowing Cambridge-Analytica to access millions of user's personal data in order to provide analytical assistance to the 2016 presidential campaigns of Trump and Ted Cruz. From quarter 2 of 2022 to quarter 2 of 2023, 46.8 user accounts per 1000 users have been breached in the US (SurfShark, 2023). Regardless of the company's intentions, the information is leaking out. It is no surprise that Americans have little faith in companies' ability to be accountable with their data. According to the Pew research center, only 5% of surveyed Americans are very confident that companies will promptly notify them if their data has been leaked and over 60% have little to no confidence at all (Pew Research Center, 2019). It's even worse when this stored data is used for predatory behavior.

Businesses employ a marketing strategy known as micro targeting: using data collected by websites such as Google or X to show the user curated ads they are more likely to interact with. Targeted ads have been used in nefarious ways. University of Arizona launched ads "that targeted poor people with the bait of upward mobility…and it worked. Between 2004 and 2014, for-profit enrollment tripled, and the industry accounts for 11% of the country's college and university students" (O'Neil, 2016, p. 70). The harsh reality is that the degrees offered by these institutions don't hold much weight when trying to find a job. The poor population is particularly vulnerable to this abuse because of criticisms that they aren't doing enough to "get ahead". O'Neil was a data scientist at a startup, when a venture capitalist debating whether to invest told her about this strategy. She later summarizes "Anywhere you find great need and ignorance, you'll find predatory ads" (O'Neil, 2016, p.70) They know what they are doing when they target social groups such as the elderly who tend not to be tech savvy or the poor who often lack opportunities for upward mobility and will jump for any chance to do so.

Once again, ANT provides a nuanced perspective on this issue. On one hand, users want to protect their privacy and give as little information about themselves as possible. On the other, they would like to "pay with their data" and have access to higher quality recommendations. Advertisers and social media companies would love more data on their users, but that comes with risks they need to manage. If they harvest too much user data, the users will be concerned and choose a different platform. Many open source and proprietary software has been made to

control the sprawl of data collectors and advertisers. Most notably, there are many free extensions on Google's own browser plug-in store that allow users to disable ads and trackers on any website they want. This allows users to reap the benefits of many websites which use advertisements to stay afloat while not having to actually view said advertisements. There are also virtual private networks which allow users to mask their IP address from nosy websites and use a proxy instead. These two tools show that users will not just stay idle and let companies milk them of their information.

Recommender systems, while providing convenience and personalization, come with concerning side effects. Their addictive design can have detrimental effects on users' productivity and well-being. The biases inherent in these systems raise ethical questions about their influence over user preferences and their potential to enable misinformation, micro targeting, and extremism. The data used to train these systems is collected, stored, and used dubiously. While these social media platforms may only answer to shareholders and advertisers, their action or inaction surrounding recommender systems has large social consequences that they cannot ignore.

References:

Anti Defamation League "The January 6 Effect: An Evolution of Hate and Extremism". Anti Defamation League. (2022). https://www.adl.org/january-6-effect-evolution-hate-and-extremism

Amnesty International. (2024, January 30). *Driven into darkness: How TikTok's "for you" feed encourages self-harm and suicidal ideation*. https://www.amnesty.org/en/documents/POL40/7350/2023/en/

Chuai, Y., Tian, H., Pröllochs, N., & Lenzini, G. (2023, July 16). *The roll-out of community notes did not reduce engagement with misinformation on Twitter*. arXiv.org. https://arxiv.org/abs/2307.07960

Dixon, S. J. (2013, July 9). *U.S. social networks who suffer from FOMO as of June 2013*. Statista. https://www.statista.com/statistics/262138/percentage-of-us-social-networks-who-suffer-from-fomo/

Dixon, S. J. (2024, February 2). *Biggest social media platforms 2024*. Statista. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Evans, Charles, and Atoosa Kasirzadeh. "User Tampering in Reinforcement Learning Recommender Systems." ArXiv.org, ArXiv, 2 Nov. 2022, https://arxiv.org/abs/2109.04083.

Evitts, Jared. "Tiktok-Addicted Students Delete App during Exams." BBC News, BBC, 4 Sept. 2022, https://www.bbc.com/news/uk-wales-62720657.

Fernandez, Miriam, and Alejandro Bellogín. “Recommender Systems and Misinformation.” Ceur, University of Madrid, 25 Sept. 2020, http://ceur-ws.org/Vol-2758/OHARS-paper3.pdf.

Friedman A., Knijnenburg B.P., Vanhecke K., Martens L., Berkovsky S. (2015) Privacy Aspects of Recommender Systems. In: Ricci F., Rokach L., Shapira B. (eds) Recommender Systems Handbook. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7637-6_19

Guo, W., Krauth, K., Jordan, M. I., & Garg, N. (2021, October 4). *The stereotyping problem in collaboratively filtered recommender systems*. arXiv.org. https://arxiv.org/abs/2106.12622

Gupta, M., Jassi, A., & Krebs, G. (2023). The association between social media use and body dysmorphic symptoms in young people. *Frontiers in Psychology*, *14*(1). https://doi.org/10.3389/fpsyg.2023.1231801

Law, J. (1992). Notes on the theory of the actor-network: Ordering, strategy, and heterogeneity. *Systems Practice*, *5*(4), 379–393. https://doi.org/10.1007/bf01059830

O’Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Penguin Random House.

Pew Research Center. (2019, November 15). *4. Americans’ attitudes and experiences with privacy policies and Laws*. Pew Research Center: Internet, Science & Tech. https://www.pewresearch.org/internet/2019/11/15/americans-attitudes-and-experiences-with-privacy-policies-and-laws/

Sparr, Matthew. “Explicit User Manipulation in Reinforcement Learning Based Recommender Systems.” ArXiv.org, Arxiv, 20 Mar. 2022, https://arxiv.org/abs/2203.10629.

Suderman, A., & Goodman, J. (2023, October 2). *Amid the capitol riot, Facebook faced its own insurrection*. AP News. https://apnews.com/article/donald-trump-technology-business-social-media-media-07124025bdbeba98a7c7b181562c3c1a

SurfShark. (2023, December 6). *U.S. Data Breach Density 2023*. Statista. https://www.statista.com/statistics/1419128/breach-density-us/

Statista. (2023, November). *Social Media Advertising - Global: Market Forecast*. Statista. https://www.statista.com/outlook/dmo/digital-advertising/social-media-advertising/worldwide

TikTok. (2024, January 24). *Privacy policy*. TikTok. https://www.tiktok.com/legal/page/us/privacy-policy/en

Twitter. (2023, March 31). *Twitter's recommendation algorithm*. Twitter. https://blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm

U.S. Department of Health and Human Services. (2022, March 22). *Drugs and the brain*. National Institutes of Health. https://nida.nih.gov/publications/drugs-brains-behavior-science-addiction/drugs-brain

VCU. (2023, March 3). *Social Media and political extremism*. L. Douglas Wilder School of Government and Public Affairs . https://onlinewilder.vcu.edu/blog/political-extremism/