Insight

Assessing the Impact of the Widespread Adoption of Algorithm-backed Content Moderation in Social Media

JUAN LONDOÑO | JANUARY 25, 2022

Executive Summary

- The use of algorithm-backed content moderation systems by social media platforms has drawn increasing skepticism in recent years, as some are concerned these platforms are responsible for polarization and erosion of users' privacy.
- While this technology has grown more sophisticated since its inception, its shortcomings have become more evident to users and critics: It is vulnerable to exploitation, it lacks a comprehension of context, and it relies heavily on a platform's good judgment.
- Despite these shortcomings, algorithms have provided users with a better web experience, allowing platforms with large userbases to present well formatted, relevant content to their users on a relatively consistent basis.
- Congress' initiatives to regulate algorithm use often fail to consider the benefits of algorithmic content moderation and could unintentionally make users worse off.

Introduction

Social media's role as a popular tool for social interaction, discussion, and argument has sparked multiple debates over how the content shared through these platforms is delivered to users. There are serious concerns over what content platforms decide to show and the process behind that decision-making. The content these platforms show their users is an essential piece of their business model, as it determines how much time users spend on their platform, how much content they consume, and what the platform's brand looks like.

In the early stages of social media, most of the content displayed was limited to what a user's social network and "followed" pages posted, usually ranked by chronological order. Platforms have continuously moved away from this strategy, instead favoring heavily curated, targeted content they believe users will enjoy more to increase user retention and screen time. To deliver this curated experience at scale, platforms now rely on automated systems backed by an algorithm to deliver the content users want to see.

These automated systems rely heavily on data collection, which has raised concern over potential privacy implications. Additionally, by pivoting into a more heavily curated experience, platforms' editorial power is more visible to users and policymakers. This, in turn, has raised concern regarding the potential dangers involved in the adoption of these systems. These concerns have been magnified following the publication of the "Facebook Files" report in *The Wall Street Journal*, and the activism of Francis Haugen, a former Facebook employee who is advocating for increased public scrutiny of social media's content moderation practices.

With skepticism toward these automated systems growing, Congress has introduced a series of bills aiming to condition or even prohibit the use of algorithms for content moderation in social media. Nevertheless, while prohibiting the use of these content moderation systems could address some of lawmakers' concerns, such action could harm not only platforms but their users, too, who could be left with a lower quality product and greater exposure to undesirable content. This insight assesses the overall impact of algorithms in social media and explains the reasoning behind its widespread implementation.

Rising Skepticism Toward Automated Systems

The introduction of bills such as the Filter Bubble Transparency Act, Justice Against Malicious Algorithms Act, the SAFE TECH Act, Civil Rights Modernization Act, and the Protecting Americans from Dangerous Algorithms Act demonstrates that lawmakers have significant concerns over algorithm-backed automated content moderation systems, usually referred to simply as "algorithms." The authors of these bills argue that the introduction of these algorithmic recommendation systems has led to an erosion of users' privacy, the spread of misinformation, further division and polarization of the country, and harmful impacts on teens' mental health.

Aside from current legislation, there have been cultural expressions of concern over algorithms in social media. The Netflix documentary *The Social Dilemma* – which had over 38 million views – heavily criticized social media platforms, particularly Facebook, for the data collection required by their algorithmic ranking systems. The documentary joined the voices of other social media critics who claim that Facebook's move to prioritize "meaningful social interaction" over time spent on the platform has instead prioritized divisive content by boosting posts with more comments and reactions. Of course, it is often heavily political and controversial content that tends to spur the most engagement.

This change in the algorithm has also made the editorial power of platforms more visible. In the "Holding Big Tech Accountable" hearing of the House Energy and Commerce Subcommittee, legislators pushed back against the platforms' power to "shape our reality," as they put it, given that the content displayed in users' timelines or news feeds is constantly subjected to the algorithm's ranking. Concern about platforms' editorial power is widespread, but the reasoning behind this concern differs between political parties. Those on the right tend to argue that platforms are removing too much content and silencing conservative opinions, while those on the left often argue that the algorithms promote the spread of misinformation, hate speech, and political extremism. These concerns have spurred the introduction of bills to regulate or prohibit the use of algorithms for content moderation to limit the moderation capabilities of platforms.

Algorithms Improve Users' Online Experience

As concern over algorithms grows, it is important to note their beneficial uses, including those that extend beyond content moderation. Algorithms are also used for technical reasons, such as recognizing a user's device to login to a platform and determining how content is displayed, such as where ads are placed on a screen or whether a website should display its mobile or desktop version. Some of the introduced bills, such as the Filter Bubble Transparency Act could – inadvertently – outlaw the use of such algorithms.

The most significant (and often overlooked) benefit of algorithm-based recommendation systems is that they help expose users to more content they are likely to enjoy. Platforms use these algorithms to actively shape content toward users' interests, prioritizing content users might find most relevant or useful. The Facebook internal documents leaked by whistleblower Francis Haugen – perhaps unintentionally – highlight the positive impact of these systems on users' platform experience. These documents revealed that in 2018, a Facebook

researcher conducted an experiment which largely disabled the ranking algorithm for 0.05 percent of users for a limited time. The experiment showed that, among these users, engagement dropped significantly, and they hid 50 percent more posts from their timelines. Of note, the study also found that ad revenue increased as users had to scroll for longer periods of time to find relevant content.

In the same way that automated systems can aid platforms in promoting interesting content, automated systems relying on algorithmic decision-making can also help to remove undesirable content in a timely manner. Content moderation has become a significant issue for larger platforms. As platforms grow, it becomes more difficult for them to moderate content as the quantity of users increase, which usually translates to a greater volume of posts that require review. Using automated systems becomes a necessity for bigger platforms, relying on them to review and flag or remove content quickly and at scale to provide a desirable service for both users and advertisers. Without the use of algorithms, platforms would be slower to remove illegal or undesirable content, such as that featuring child sexual abuse or violent and sexually explicit material.

Algorithms' Biggest Shortcomings

Despite their benefits, these automated systems have various shortcomings that have become more evident as the technology matures. A shortcoming that has drawn significant criticism from both sides of the aisle is the algorithms' dependency on the good judgment of the companies that employ them. Specifically, an algorithm's administrators decide which variables comprise it and the influence it exerts. One of the biggest concerns regarding the design and training of algorithms is "algorithmic bias," which refers to the situations in which the variable or source data behind an algorithm can lead to bias it against certain groups regarding their ethnicity, political party, sexual preference, gender, or race. Platform experimentation with prioritizing or deprioritizing certain controversial topics can have unintended consequences for users if they are not satisfied with the moderation outcomes.

Another major shortcoming of reliance on algorithmic systems is that they often lack the capacity to understand the context in which content is posted. This has been particularly evident with the enforcement of copyright regulation online. Due to the strict nature of the Digital Millennium Copyright Act (DMCA), platforms have relied heavily on algorithms to identify and remove potentially infringing media content in a timely manner. But these automated systems have resulted in cases in which content is taken down due to unintentional or accidental reproduction of copyrighted material, such as music being played by a loud car or a store speaker. Because these systems use algorithms reliant on "objective" variables to make decisions, they often prove inept at understanding the context in which content is posted.

In recent years, there has been increased attention to how automated systems can be exploited by malicious actors. As algorithm technology has matured and the variables that drive the algorithm have become more predictable – due to either platforms publication of algorithm rules or users' familiarity – malicious actors such as those posting spam or fraudulent content can trick the algorithm into boosting their reach. Automated systems can also be exploited by those looking to censor or shut down speech, which has happened in cases where malicious actors abused copyright regulations to shut down content creators through copyright reports, and when police officers have reproduced copyrighted content—for example, playing copyrighted music—while being recorded by civilians to trigger social media algorithms to kick the video from the platform.

The Dangers of Current Proposals on Algorithms

Current initiatives in Congress propose various approaches to regulate the use of automated systems by social

media websites. Some bills, such as the Filter Bubble Transparency Act, would require platforms to disclose when they use information users did not actively provide to rank the content displayed on a website. It would also require platforms to provide an algorithm-free ranking alternative. Meanwhile, the Justice Against Malicious Algorithms Act would strip the protection provided under Section 230 of the Communications Decency Act for social media websites that use algorithms to promote content. This could lead platforms to reconsider the use of algorithm-based systems, as the potential risks of being held liable for third-party content could possibly outweigh the potential benefits of using them. Other bills including the SAFE Tech Act, the Civil Rights Modernization Act, and the Protecting Americans from Dangerous Algorithms Act would also remove Section 230 protections for certain uses of algorithms.

These types of bills would disincentivize platforms' use of automated systems, particularly as the potential risks associated with lifting Section 230 protections could make them liable for content posted by third-party actors. This could prove seriously harmful, because as mentioned before, algorithms are present at various levels in online platforms with various applications. Most of these bills would target algorithms present in content moderation to limit users' exposure to harmful or undesirable content. But due to the various and essential uses of algorithms, they could also seriously impact the broad majority of user-friendly applications. There have been various complaints regarding the wording of these bills, as they often employ terms as "secret" or "obscure" algorithms, which are abstract and ambiguous and could lead to difficulties when enforcing these laws. Critics claim that, as currently defined, most of these bills would unintentionally target consumer-benefitting algorithms such as those modifying content based on location or device, which would provide consumers with irrelevant or poorly formatted content. Additionally, some of these bills could prove counterproductive to their stated objectives, as platforms would be hesitant to use algorithms to remove misinformation or illegal content if they face higher legal liabilities for their use due to the removal of Section 230 protections.

Conclusion

Online platforms have widely employed algorithm-backed automated systems for content moderation to provide the best possible experience to their massive userbases. These systems can quickly consider a vast number of variables and yield relevant and appropriately formatted content to users at scale that would be impossible without the use of automated systems. Nonetheless, as the technology has matured, some of its shortcomings have become more evident, such as its dependence on platforms' good judgment, its frequent inability to comprehend context, and its exploitability. Congress' initiatives for regulating algorithm use tend to ignore the benefits of algorithmic content moderation. Additionally, the bills' use of ambiguous definitions could lead to unintended harm to consumers. While flawed, automated systems have proven to be a net-positive for users that benefit from more relevant and desirable content.