Introduction

As technology and society continue to merge, engineering practice increasingly involves complex technosocial systems whose impacts can't always be easily traced back to designs. This manifests in one such engineering practice where engineers are tasked with creating systems that maximize user engagement for social media platforms. These systems are effective at optimizing a user's attention and generating profit, but can also negatively impact certain users' mental health and have broader adverse effects on society. An ethical dilemma occurs when engineers try to determine whether or not to take action. The difficulty of mitigating these problems is also heightened by the natural complexity of societal impacts.

Background

Machine learning models are used in a variety of ways in social networking apps, like filtering out offensive content or ranking search results. However, models designed to track and increase engagement usually automate content ranking and recommendations (Dunn & Dunn, 2018). Content ranking models across platforms are more similar than they are different. Generally, they work like this: For every post on a user's feed, the algorithm receives feedback on whether or how the user interacted with that post. Then, the algorithm uses this data to rank available content according to how likely it is that the present user will engage with it. Different applications may have different measures for how engagement is ranked. For example, Facebook uses the weighted average of different user actions like comments, reshares, reactions, or likes to measure engagement, and YouTube uses watch time to measure engagement (Narayanan, 2023). How algorithms optimize engagement can differ between signal and computational techniques. But generally, they use three main types of signals to indicate similarities between users: network, behavior, and demographics. Network is used more by social networking apps like Facebook, and dictates two users as similar if they comment, follow, or subscribe to the same people. Behavior categorizes users as similar if they interact with the same posts. Demographics is usually only used if the user has very little behavior data, like when they first join the application. The algorithm also breaks down which posts are similar using signals like content and behavior. Using content to categorize posts by different attributes and behavior to group posts that have interactions from similar users. The behavioral record is the driving force for the recommendation algorithm. TikTok, for example, has interaction records on over half a million videos for a single user. (Narayanan, 2023).

Research shows that content ranking algorithms increase user retention. In a research article posted by Science, researchers found that a chronologically ranked feed dramatically reduced users' time and engagement on platforms like Facebook and Instagram. On Facebook, participants on the chronological feed liked only 3.7% of the content they were exposed to, which is less than the participants on the regular feed who liked 6.7% of the content. Participants on the regular feed also commented on more content present in their feed, 1.5% compared to 0.8% (Guess et al., 2023).

Researchers assert that content ranking models have negative effects on mental health, with young adults and teenagers being the most vulnerable. Studies have shown that heavy and prolonged use of social media has contributed to worsening symptoms of depression and anxiety (Naslund et al., 2020). Some researchers have connected this with content ranking because users who engage with sad or depressing content can spiral into increasingly consuming more negative material (Hao, 2022). Research has shown that the tools used to increase engagement can cause excessive screen time and limit people's ability to function. Researchers are divided on social media's addictiveness. With some cautioning that clarifying it as an addiction may be misrepresenting the term, and others saying it does have similar symptoms to addictions like gambling. Researchers assert that it offers an instant dopamine hit that creates a powerful urge for users to keep engaging. They also assert that mental illness often underlies addictive habits. On the other hand, researchers note that social media can improve mental health, helping some people build connections and community (Naslund et al., 2020). It should also be noted that all these effects are very new, and scientists still continue to study social media's effects on society. However, they do assert that social media's harmful effects cannot be denied even if causation can't absolutely be proven (Naslund et al., 2020).

Other researchers are concerned about "content bubbles" causing social division and extremism. Some studies show that models made to maximize engagement increase polarization (Hao 2022).

Ethical Reasoning

I will be using reflexive principilism to analyze my case. Reflexive principilism will allow me to take a flexible approach as my case has various complexities that cannot be resolved by other ethical frameworks. My reasoning will be highly based on judgment, with some weight on actual actions, so fixed rules would not work for my case. I wanted to include other ethical ideas like deontology, consequentialism, and virtue ethics in my analysis. Additionally, I thought the principle of autonomy would be especially relevant for my case.

Premise and Stakeholders

Platform industries commission engineers to create machine-learning models to keep user retention. However, models designed to maximize engagement have been shown to have consequences on societal discourse and users' mental health. This study examines whether harmful effects should be addressed at the cost of business or sometimes professional standing, or if they can't be totally proven and are outside an engineer's purview.

Stakeholder #1: Management at Social Media Companies

Social media is a fiercely competitive industry that uses different features, content, and algorithms to vie for a singular, limited resource: users' attention. Social media platforms offer free services in exchange for users' data, which they then sell to advertisers to make money. So, users are effectively paying for a platform's services with their information and attention; this means social media companies can only make money if users are actually on their application.

So if users lose interest and move on to other applications, platforms effectively lose business (Huo, 2022). Social media companies know that users who are engaged are more likely to return and generate more ad revenue for the platform. That is why social media platforms commission engineers to develop tools to increase users' engagement.

The industry is dominated by leading companies like Meta, TikTok, and YouTube. Machine learning models are a major way they stay ahead in the industry. Companies that do not have sophisticated models would be unable to compete with other platforms that do. So removing these models would result in a major loss of revenue.

Some companies also do not accept that their algorithms are solely responsible. Claiming they have addressed concerns raised about users' safety and privacy, and issues are misrepresented by misinformed opinions. Asserting that they have done research on its effects and addressed them accordingly (Duffy, 2025). Facebook, for example, created a team to address content ranking issues (Hao, 2022).

Stakeholder #2: Social Media Users

Users of these applications include all types of varying demographics. In 2025, about 5.45 billion people were estimated to use social media around the world (Elad, 2025). Users vary in their perspectives concerning issues about mental health and societal discourse. Some users like how their content is personalized to their tastes. Users are able to open up an application and easily find content that interests them. Other users raise concerns about the negative impacts of social media algorithms.

Researchers: This includes psychologists and scientists who have studied the effects of social media on society, especially concerning the effects of content ranking models. There is continual research being done on its effects, which has been hindered by the fact that social media companies do not give them access to any of their data and have not released any internal research done on content ranking models negative effects.

Lawmakers: Several hearings have been held by Congress to address several issues concerning social media. With some people asserting that the social division and extremism created by models break down democracy. Lawmakers are politically divided on whether scanning models for misinformation is censorship. However, most agree on issues like protecting users' mental health, especially for young adults and teens.

Parents: Parents of young teens and children have raised concerns over algorithms' effects. Asserting it causes addiction and worsens symptoms of depression. Several lawsuits have been filed against social media companies by parents and school districts alleging that social media addictions harmful effects (Duffy, 2025).

Young adults and teens: Teens and young adults have grown up on social media, and it is a very significant part of their lives. Teens themselves have admitted that social media has made their mental health worse, with 30% of teen girls saying Instagram has made them feel worse about their bodies (Hao, 2022). However, even though some people acknowledge its negative impacts on focus and mental health, overusage is still normalized. Some teens assert a lack of transportation and fewer communal spaces as a reason why they spend more time online.

Younger generations' hobbies and community can also be online, which can contribute to greater usage.

Non-English speaking users: The risk of algorithm influence can be greater for non-English speaking regions. The algorithms aren't trained for these regions as well, so the risk of polarization is increased. Algorithms have been seen promoting extremist groups; 64% of joins can be attributed to recommendation tools (Hao, 2022).

Stakeholder #3: Engineers

Engineers are also divided on this issue. Many expressed concerns over discrimination, the spread of harmful content, and reduced critical thinking. Engineers have claimed concerns have been largely ignored by management, and models that decrease engagement significantly are thrown out (Haos, 2022). In reaction to this, some people have risked their professional standing to spread awareness of harmful effects. For example, Frances Haugen collected and released documents about content ranking algorithms harmful effects before quitting Facebook in 2021 (Haos, 2022). Others who have raised concerns have remained anonymous in fear of losing their livelihoods. When some engineers state that addressing issues like misinformation can be complicated. There is a large debate on what is considered misinformation. For example, anti-vaccination groups consider removing their content to decrease misinformation as censorship. Trying to decide what misinformation is can be politically divisive. Even if misinformation is penalized, it still persists. Algorithms must be trained on what information to penalize, but misinformation is constantly evolving. You avoid a penalty by simply changing your wording. Engineers have tried to build models based on fairness, but there are different definitions of fairness that don't coexist (Hao, 2022). Engineers also claim that the algorithm isn't the enemy and personalization is a major thing that makes platforms entertaining. Emphasizing transparency about how algorithms work and teaching media literacy as important factors in combating these issues (Guess et al., 2023).

Ethical Analysis

Respect for autonomy: Can be defined as protecting people's ability to meaningfully choose what actions they take without being manipulated or deceived in any way to make certain decisions.

In the context of this case study: Prompts the question: Do users willingly choose to engage with applications? Many users have expressed that they enjoy the social connection, self-expression, information, and entertainment that social media gives them. So users do willingly choose what applications they use and what content they seek out. However, there is evidence to support that algorithms have a significant impact on users' experience. Algorithms can amplify content that elicits an emotional reaction from users in order to increase engagement. It can also be argued that algorithms optimize engagement too much to the point where users display addictive behavior. Lack of transparency for how algorithms work means users cannot properly consent to their influence. This is especially the case for young adults and teens as studies have shown they are more vulnerable to their influence (Duffy, 2025). Users also

have no ability to control how they are affected by these systems. They cannot change content personalizations and do not have the ability to control the societal effects caused by them. It can also be argued that allowing false or misleading information to manifest online can distort users' view of reality. Since social media can be a source of information, it can inhibit a person's ability to make informed decisions if they are constantly fed false information on their feed.

Companies argue that content on individual users' feeds is based on their actions, and so the content that is provided is something users are seeking out. This can also apply to misleading information, implying that if it is on their feed, it was something they already wanted to see. However, they continue to fail to have transparency on the effects of their algorithms.

Nonmaleficence: Is defined as to avoid causing harm, including the mitigation and prevention of harm. Must take responsibility for negative side effects.

In the context of this case: Engineers have the responsibility to address issues technically and be transparent about the effects of their designs. The harms couldn't have been foreseen by engineers ahead of time, as techno-social systems often have unintended social consequences. However, they do have a duty to address these issues when they manifest. There have been teams of engineers put together in order to mitigate these issues, which have displayed a serious commitment to solving these issues. However, a lack of support from management can be an obstacle to putting necessary systems into place. Facebook, for example, does not make passing the ethical testing put forth by their own team of engineers a requirement for released models (Hao, 2022). Management's lack of urgency in fixing these issues is a majorly neglect of their responsibility to combat negative side effects. There is evidence to support that they knew from their own research of the problems with optizmizing engagement, but did not act on any findings as any resolution would decrease engagement (Hao, 2022).

It should be noted that the mitigation and prevention of harm is difficult to address. Causation is difficult to identify with societal harm. Especially with claims like algorithms undermine democracy. Algorithms may not be able to be sufficiently reworked to address these issues.

Beneficence: Means to actively work to promote people's well-being and create positive outcomes.

In the context of case study: Algorithms do effectively entertain people. The long period people spend online can be attributed to how much they enjoy the platform. The idea behind them was to make it easier for people to find things they enjoy. Algorithms can introduce them to new communities, information, and people they never could have found before. Part of the idea in creating them was to create happiness. However, it doesn't mean it promotes their well-being. Social media can disrupt people's focus and cause them to put off their responsibilities. So it creates some positive effects, but may not promote people's well-being.

Justice: Means fair distribution of benefits and risk. Protections, especially in the case of vulnerable individuals.

In the context of case study: The risk algorithms cause is not equal among everyone. Most significantly in our case, minors and young adults are more likely to be affected by

algorithms influence. Teens', especially girls', mental health is negatively impacted by algorithms as they are shown more content that contributes to their symptoms (Hao, 2022). Algorithms reflect the discrimination of society at large. Marginalized communities have reported reduced visibility on applications due to algorithm bias. Creators have also pointed out discriminatory practices in which content is removed by the algorithm, citing a post that was removed for nudity for a black plus-sized creator vs posts of thin white women in a similar post still up on platform (Willcox, 2025). Definitions of what misinformation is complicate how algorithms can fairly remove false information. For example, organizations that promote anti-vaccination campaigns claim their content is unfairly taken down.

Reflections on Ethical Tensions

The main tensions of my case can be best explained by autonomy and nonmaleficence, and will ultimately have the most weight in my decision. Autonomy establishes that the algorithm's influence on users is real and disrupts people's decision-making. It also shows how design practice and nontransparency can limit users' understanding of how systems affect them. Nonmaleficence shows the ethical complexities in resolving this issue. Additionally, justice points out the ethical complications that some resolutions have. Beneficence has some points promoting algorithms and will help give a more nuanced conclusion.

Principles like beneficence and non-maleficence will conflict; they are naturally opposites in this case, but non-maleficence will have more weight because harms cannot be justified. Justice will conflict with non-maleficence as well because unfairness is an obstacle in finding a resolution that prevents harm. Removing content to prevent misinformation creates outcry over discrimination and censorship. Beneficence and autonomy conflict because different interpretations can be taken from excessive usage of applications. Engineers' main problem in resolving these issues is complicated by their duty to the organization they work for, themselves, and society. Content personalization is a major way organizations make money, and removing algorithms would limit their ability to grow. Pressure from management and the risk of professional standing cause engineers to hesitate to voice their concerns. Personally, engineers need a job and money like everyone else, and raising concerns could put their livelihoods at risk. However, engineers also have a duty to society to prevent and mitigate harm if it is identified.

Possible solutions:

1. **Maintain current design practices:** Algorithm will continue to prioritize engagement and will rely on users to practice media literacy and self-regulation. Puts responsibility in users' hands to critically evaluate content they consume, and to limit the amount of time they spend on a platform.
2. **Practice transparency, promote media literacy, and introduce user controls:** Companies and engineers will be open about designs and how they affect users. This may mean releasing more data to researchers so they can get more conclusive findings on algorithms effects. Make it a part of the platform's mission to encourage users to analyze and evaluate the content they consume as well as promote friendly media usage. This could also mean creating controls for turning

off certain parts of the recommendation algorithm, so users can choose whether or not they want to receive content recommended by models.

3. **Redesigning algorithms, along with transparency, media literacy, and user controls:** Engineers would prioritize reducing harmful and misleading content even at the expense of engagement. This means engineers and companies would take more technical actions to reduce these issues. These steps may include taking down or deprioritizing content that is classified as misinformation, or displaying on posts that the content has been classified as misinformation, along with a source. Filtering content for quality may be another way to do this. At the very least, algorithms should be perfected so problems aren't magnified for non-English speaking regions and violent extremist groups should be banned from applications.

Justified Resolution:

The third option: redesigning the algorithm, along with transparency, media literacy, and user controls, aligns with my analysis the best. It does the most to mitigate the problem while giving users the ability to meaningfully consent and control how systems affect them. While it may reduce engagement and profitability, I believe the application will still be engaging. Social media is so ingrained into society that I don't believe changes would have a significant harm on revenue, and companies make more than enough money already. Engineers should prioritize public welfare even when doing so conflicts with corporate goals. Engineers should also consider all aspects of impacts, even if that means considering the outermost social impacts. Though some impacts may lie outside an engineer's purview, I think it's important to consider all societal impacts as techno-social systems become more and more dominant in society.

Limitations and Ongoing Responsibilities

This resolution does not eliminate all ethical concerns. Doing the research for this case study made me seriously question whether it was possible to change algorithms to address the issue. Especially concerning societal impacts, as forms of communication that prioritize engagement naturally undermine institutions and public discourse. However, I also don't think there is a big enough push in the industry to try and find a technical solution. I would highlight that transparency and education are the most important factors in this resolution. More people should understand how algorithms affect them and know how to practice media literacy. As techno-social systems become more and more prevalent there may be additional impacts from content ranking algorithms. So, engineers should monitor effects of systems on society.

Sources:

Duffy, C. (2025, November 26). *Lawsuit alleges social media giants buried their own research on Teen Mental Health Harms | CNN business*. CNN.

https://www.cnn.com/2025/11/25/tech/social-media-youth-mental-health-lawsuit-meta-tiktok-snap-youtube

Dunn, J., & Dunn, J. (2018, June 28). *Introducing Fblearner Flow: Facebook's Ai Backbone*. Engineering at Meta. https://engineering.fb.com/2016/05/09/core-infra/introducing-fblearner-flow-facebook-s-ai-backbone/

Elad, B. (2025, December 15). *AI in Social Media Tools Statistics 2025: Uncover what's shaping the future*. SQ Magazine. https://sqmagazine.co.uk/ai-in-social-media-tools-statistics/

Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Kennedy, E., Kim, Y. M., Lazer, D., Moehler, D., Nyhan, B., Rivera, C. V., Settle, J., Thomas, D. R., … Tucker, J. A. (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, *381*(6656), 398–404. https://doi.org/10.1126/science.abp9364

Hao, K. (2022, June 29). *The facebook whistleblower says its algorithms are dangerous. here's why.* MIT Technology Review. https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/

Huo, Emily. Moral Arguments on Engineers' Actions when Designing Social Media for Addiction. University of Virginia, School of Engineering and Applied Science, BS (Bachelor of Science), 2022-12-18, https://doi.org/10.18130/t45j-0834.

Narayanan, A. (2023, March 9). *Understanding social media recommendation algorithms*. Knight First Amendment Institute. https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms

Naslund, J. A., Bondre, A., Torous, J., & Aschbrenner, K. A. (2020, September). *Social Media and Mental Health: Benefits, risks, and opportunities for research and Practice*. Journal of technology in behavioral science. https://pmc.ncbi.nlm.nih.gov/articles/PMC7785056/

Willcox, M. (2025, December 14). *Algorithmic agency and "Fighting back" against discriminatory Instagram content moderation: #iwanttoseenyome*. Frontiers. https://www.frontiersin.org/journals/communication/articles/10.3389/fcomm.2024.1385869/full