Rabhya Mehrotra

An Alternative Approach to Content Moderation: Expanding News Feeds Through Randomization

It is no secret that the content posted on Facebook can both empower and harm communities all around the world. As extreme (and illegal) posts have proliferated, Facebook and other social media platforms have been subject to increasing pressure by governments and civil society alike to address these issues through content moderation. Current attempts at content moderation, such as taking down specific posts or regulating algorithms, have not sufficiently addressed the problem. I will therefore propose a third attempt at content moderation: expanding what users see through randomization.

I will outline my proposal as follows. First, I will describe Facebook's News Feed, and how Facebook's algorithms decide what posts to show. Next, I will look at current efforts at content moderation and the challenges they face. I will then propose an alternative approach: in addition to limiting what users see, I will propose that Facebook expand what users see. Using the example of BitTorrent, a peer to peer network that exchanges information, I will show how Facebook can systematically incorporate randomness into its algorithm. I will then discuss why this approach to content moderation may affect user's perspectives, and why such an approach is financially attractive for Facebook.

An Introduction to Facebook's News Feed:

Facebook's News Feed is the platform's crucial product. Out of the 50 minutes the average user spends on Facebook each day, most of that time is spent on the News Feed rather than any other part of the website (NYT). A News Feed is composed of various posts from a user's friends, groups, liked pages, and of course, sponsored content. How does Facebook decide what is on a user's News Feed? This question can be answered by examining Facebook's promise to users and advertisers. Users remain on Facebook because it promises an engaging platform, filled with content that users would like to see. Advertisers, meanwhile, remain in business with Facebook because it promises them the user's attention. Companies pay Facebook for precious slots in a user's News Feed, and those slots become more precious when Facebook can target ads specifically to users. These two goals are mutually beneficial: the more time users spend on a News Feed, then, the more time Facebook can "learn more about its users — their habits and interests — and thus better target its ads" (NYT). It is thus crucial for Facebook's News Feed to hook users on for as long as possible, and to encourage engagement so it can learn what its users like and dislike.

In 2018, Facebook noticed that its user engagement was decreasing. This was a fundamental threat to their business model, and the company scrambled to find a solution. Later that year, Mark Zuckerberg released a new kind of measurement: Meaningful Social Impact, or MSI. MSI was an elaborate numerical points system used for each post, whereby Facebook's algorithms could understand how the post fared with its intended audience. Each like was worth one point; a "non-significant" comment was worth fifteen; a "significant" comment was worth thirty; and so on (Hagey). Posts were also judged based on *who* interacted with them: if a stranger engaged with the post as opposed to a close friend, their engagements were devalued in the points system. MSI marked a shift in Facebook's News Feed: its algorithms were now meant to maximize MSI above all, making sure that posts were matched with users who would engage with them the most.

To maximize MSI, algorithms aim to show users posts they either strongly agree with or strongly disagree with because both prompt strong reactions. This was helpful for increasing Facebook's short-term ad revenue, but detrimental to content moderation efforts on Facebook. First, when users only receive posts and articles that fit into their existing worldview, they enter into echo chambers, which only deepen the ties between maximizing user engagement and spreading polarizing content. Second, the existence of echo chambers (and high ranking of inflammatory content within them) means that it's in the best interest of various actors – whether political parties or news organizations – to post extreme content. This has tangibly affected what kind of messaging political entities engaging in. In Poland, for example, one political party estimated that its posts shifted from 50% positive – 50% negative to 20% positive – 80% negative. An internal Facebook report found that this shift came "explicitly as a function of the change to the algorithm" (Hagey). Given that Facebook's platform is a place for anyone to post, and problematic posts are often more popular (and therefore can inflict more harm), what can the site do?

Current Approaches to Content Moderation:

One approach to content moderation has been flagging and removing specific posts and messages that are not allowed. Facebook has a list of Community Standards, which regulate what can and cannot be posted on the site. To enforce these standards, it employs over 30,000 human moderators and extensive AI capabilities, which flag down and check problematic content (Douek). These efforts, however, are not enough to deal with the sheer volume of the problem: "over 2.5 billion pieces of content are shared on [Facebook's] platform every day" (Douek). Moreover, there are numerous issues with cultural content and language. What might be considered a reclaimed term in one culture could be offensive in another: the term *queer* is now used by members of the LGBT+ community. The word originated as a slur, however, making it difficult for moderators to understand when its usage is appropriate and when it is not. Most importantly, these efforts do nothing to address the *incentive to create* extreme and false posts.

Another approach to content moderation has been algorithmic regulation. It attempts to address the incentive to post extreme content: instead of looking at *what* posts are left up on the website, algorithmic regulation attempts to study *who* gets to see them and why. For example, posts that are typically "borderline" to being banned are often the most popular because of their inflammatory content. For "borderline" content in each of its harmful categories, Facebook works to "distribute that content less" to reduce the incentive to post such content" (Douek). Some politicians, such as Donald Trump, often complain that they have been 'shadow banned' by platforms like Facebook, meaning that their content is intentionally distributed less. Indeed, Facebook's algorithms perform the bulk of its work in content moderation: "Facebook is increasingly relying not on the blunter content moderation tools of removing posts or pages, but on the subtler tools of limiting their reach and exposure.

While regulating algorithms is a promising field, there are several concerns. First, Facebook "has famously jealously guarded" its algorithms, and for good reason: Facebook's main product, after all, is its curation of content that keeps users hooked (Douek). Facebook is constantly competing with other social media platforms to keep users, because part of its value lies in its access to widespread networks. Furthermore, Facebook may want to prevent public access to its algorithms because bad actors may take advantage of them. Algorithmic design can teach extremist organizations, for example, how to micro-target potential recruits and maximize public access to violent content.

Second, the sheer amount of algorithms used by Facebook may make the whole enterprise extremely complicated. "Facebook decides how to target ads and rank content based on hundreds, perhaps thousands, of algorithms," meaning that there is no one algorithm that could be tweaked. Moreover, the effect of these algorithms combined is usually different from their individual actions: "some of those algorithms tease out a user's preferences and boost that kind of content up the user's news feed. Others are for detecting specific types of bad content, like nudity, spam, or clickbait headlines, and deleting or pushing them down the feed" (Hagey). It remains difficult for regulatory agencies to both identify particular algorithms and meaningfully understand how regulation would affect the entire system of algorithms.

Finally, and perhaps most importantly, Facebook has no business incentive to regulate its algorithms. If Facebook cares about profit – and by extension, maximizing user engagement – it will never voluntarily 'worsen' its own algorithm and make the platform less engaging. This has been shown in the past: in 2019, data scientists found that limiting the spread of oft-shared posts about civic and health information – even when the algorithm guessed that a user would engage with it – "helped reduce the proliferation of false content." Yet when they proposed adjusting the algorithm in a broader context, Mark Zuckerberg refused: "He didn't want to pursue it if it reduced user engagement" (Hagey).

A Possible Solution: Adding Randomization

I therefore propose a middle solution to the growing problem of extreme content. In addition to letting Facebook restrict what content users see – either by taking down posts or using algorithms – I propose Facebook *expand* what users see adding randomness into users' feeds. User's feeds are highly scripted, based on large amounts of data on the user themself and their past engagements on the site. Users can try to randomize their feeds, but these tools are rather haphazard: there is Chrome Extension called 'Go Rando,' for example, that will randomly change a user's reactions to posts. Yet there is no systematic method of randomization. Of course, we are not aware if Facebook's algorithms currently use randomization when showing posts. However, we do know that Facebook aims to show posts with high MSI values, which means my solution is distinctive from current efforts. This is because I propose that while Facebook may keep its algorithms that maximise user engagement, it ought to consistently throw random posts and articles that have a low predicted MSI score. My solution is therefore a complement, *not a supplement,* to existing content moderation efforts.

I will outline my solution in the following order. First, I will explain why incorporating randomness is possible, because there is precedent for this kind of addition to knowledge-sharing algorithms. The BitTorrent file sharing system, for example, is a peer to peer network system that exchanges information. BitTorrent uses 'optimistic unchoking,' or occasionally randomly picks its peers. I will explain why Facebook is analogous to BitTorrent's game and could therefore also incorporate randomness. I will then argue that this randomness has the potential to reduce polarization – and why that is both financially beneficial for Facebook and a creative solution to content moderation.

BitTorrent File Sharing + Cooperative Behavior

The BitTorrent file sharing system is a peer to peer network in which users engage with each other. BitTorrent's design is fairly simple: to share files, there is one agent called the 'seed,' which already has the entire file. There are other multiple agents called 'leechers,' who are all trying to download this file. The seed breaks up the file into multiple smaller pieces, which the respective leechers begin to download. Whenever one leecher downloads a piece of a file, it uploads that information to a tracker, so other leechers can access its downloaded piece. The way in which pieces are selected is crucial: one must not "end up in a situation where every peer has all the pieces that are currently available and none of the missing ones" (Johnsen et al). To address this problem, the leechers have preference methods for selecting pieces. They each start off by downloading a random piece, and then look for the 'rarest' piece, i.e the piece which is on the least amount of *other* leeches. After downloading the rarest pieces, they turn to the more common pieces, until they have finished downloading the file (Jonhsen et al).

How do peers select each other? In the game of peer selection, each peer is a player. (Here I will refer to the actors as players, but their goal as 'peer selection.') The action space for each player is the same: cooperate (share files) or defect (refuse to share). A player wants to maximize their gains (and therefore minimize their losses). Downloading a file piece increases the player's value. However, uploading a file piece to others can cost a player money and energy. This leads to players who only take and never give: called 'free-riders,' they undermine the whole enterprise. Players determine who is a free-rider by comparing their uploading and downloading rates with the other player (University of Iowa). The process of peer-selection is a prisoner's dilemma between every pair of players. If they played once, each would have a dominant strategy (meaning it was preferable regardless of the other player's move) to defect, or not share its piece of the file. However, it is advantageous for both players to cooperate because they could work towards achieving the larger goal of downloading a full file. Furthermore, players share multiple pieces of files with each other. The game of peer selection is therefore a repeated prisoner's dilemma. In a repeated prisoner's dilemma, it is also advantageous for both players to cooperate over the long-term.

BitTorrent incentivizes cooperation by using a process of ‘choking.’ A player wants to select other players who are cooperative, or willing to share their files. To ensure that other players will share, a player can cut off others who are unwilling to share their pieces. It does so by not letting them download its own pieces in the future, in a process called ‘choking.’ If there is an exchange between player A and player B and player B defects (i.e does not share its piece of the file), player A can ‘choke’ player B. Once player A chokes player B, it is assumed that A will continuously choke B (i.e refuse to share any more pieces of its file), and both will therefore permanently ‘choke’ each other. Choking helps foster cooperation: each player is incentivized to share their file pieces rather than be a free-rider, because they know that defecting will be punished. Players therefore operate on a ‘tit for tat’ premise: cooperate first, and then mirror the other peer’s actions.

While choking is usually permanent, BitTorrent has an “optimistic unchoking” element in its peer selection. Occasionally, a “client uses a part of its available bandwidth for sending data to *random* peers” (Savolainen, emphasis mine). The player is not evaluating whether the other player is also reciprocating (i.e whether it is a free rider or not); it is sending bits of files anyways. Optimistic unchoking has several benefits. First, it helps allow new players to participate in the process: when they do not have any uploading or downloading rates, other players don’t know if they are cooperative yet or not. Second, it helps the old player find a potentially better partner player than the one it had before. BitTorrent therefore exemplifies a network system in which sometimes players select each other randomly.

Facebook’s Game

Before describing how BitTorrent could apply to Facebook, I will take some time to describe the status quo. BitTorrent's players are identical, and therefore have the same incentives and same options. Facebook's players do not. In Facebook’s current game, one player is Facebook’s algorithms, adding posts to the user’s News Feed. The other player is the user, who interacts with each post on the News Feed. Each player has their own action space. Facebook’s algorithms have one option: show the user posts with maximum MSI. The user has two options: spend more time on the website (cooperate) or spend less time on the website (defect).

I will explain this action space for the user by looking at another alternative: engage, don’t engage. When the user is on the website, it can engage or not engage with posts. While this may seem like a reasonable action space, it is not because both options are still technically beneficial for (and therefore cooperating with) the algorithms player. When a user player engages, Facebooks’ algorithm player knows its MSI scores are accurate. When a user player doesn't engage, Facebook's algorithm player knows that it must choke those posts and find better ones. In both cases, Facebook’s algorithm player is simply improving its knowledge, and there is no cost to Facebook’s algorithm player if the user engages or does not engage. So why would Facebook’s algorithm player care about providing posts with a high MSI value at all?

In reality, users won’t be willing to waste their time on the platform. If many of the posts on a Facebook News Feed are dull, the user may choose to not engage with them. However, repeatedly not-engaging with posts may decrease the user's desire to spend time on the website. This is especially true when a user may begin to think a competitor social network could provide a better product. Facebook's algorithm player cares about providing posts with a high MSI value because it wants a user's attention. While the user may occasionally not engage with posts, this would not be a concern to the algorithms player. The algorithm player would only feel a loss when the user chooses to spend less time on the site because it has become boring, or a waste of its time. Therefore, the correct defect option for a user player is spending less time on the website.

It's important to understand each player's desired outcome. The Facebook algorithms player wants to maximize its gains and minimize its losses. It therefore wants to select posts that prompt the user to cooperate, or engage with. When the user cooperates, it spends longer on the website which is important for Facebook's advertisers, therefore providing a 'gain.' A post that the user does not cooperate with, however, represents a loss. The more bad posts a user receives, the less likely it is to spend time on the site. In BitTorrent, a player tries to find another cooperative player by comparing their uploading and downloading rates. Here, the Facebook algorithms player tracks the MSI values of posts for different people to see how likely each post would be to inspire engagement. The algorithm then selects a post for a user based on how high the post's expected MSI is.

The user player, meanwhile, wants to maximize its gains and minimize its losses. A user therefore wants a News Feed that is interesting and engaging, because it will therefore be valuable to spend time on the website ('a gain'). A user does not want to waste their time, especially when there is an opportunity cost of going to another superior social networking platform (loss). When the user receives the post, their action space has two options: decide to stay on the website (engage), or decide to spend less time on the website (defect). A user wants to engage with posts it likes, because it knows the algorithm will provide more of those posts in the future. It is also beneficial for the user to not engage with posts it dislikes (because the algorithms player will 'choke' similar posts in the future), but only a few times. If a user repeatedly does not engage with posts, this may indicate that they are bored with the site and feel they are wasting their time. Repeated non-engagement may lead a user to spend less time on the site.

Unlike the prisoner's dilemma, neither player has a dominant strategy of defecting. In the case of the Facebook algorithms player, it always wants to cooperate by sending the highest value MSI post, because that will keep the user engaged. In the case of the user player, it does not have a dominant strategy: its desired strategy depends on the Facebook algorithms player. If the Facebook algorithms player cooperates (i.e throws a good post), it will also wish to cooperate (engage with the post). If the Facebook algorithms player repeatedly defects (i.e throws a bad post), it will also wish to defect. The user mostly plays a tit-for-tat game, then, mirroring the other player, with one exception. There are times where the Facebook algorithm may incorrectly attempt to cooperate with the user (i.e by sending the user a post it thinks the user would like, but the user actually dislikes). In this case, the user would wish to defect.

Applying Optimistic Unchoking to Facebook

How might we address this exception to the tit-for-tat game, where a Facebook algorithm incorrectly cooperates with a user (i.e sends a post with a high MSI score that the user doesn't engage with)? Although the games of Facebook and BitTorrent are slightly different, there are lessons from BitTorrent that can still be applied. In particular, I propose that Facebook could introduce an optimistic unchoking into its algorithm player's decision-making. This would change the action space of the algorithm player. It can either show a post that is chosen at random from the set of all available maximum-MSI items, or display a post that is chosen uniformly at random from the set of all available items with c*MSI scores, where c = .5. The number c is a positive constant that can exist in a range (0,1), but for the purpose of this paper I will assume there is only one c value. The user, meanwhile, has the same action space.

Just like in BitTorrent, optimistic unchoking could help both the Facebook algorithm player and the user player. In the case of the Facebook algorithm player, its expected MSI values may not align with the user's preferences. This could be for a variety of reasons: the user may have totally random preferences that are difficult to predict, the past posts that the user has reacted to were not comprehensive, and so on. When the Facebook algorithm player engages in optimistic unchoking, it could help find better posts for the user than it had in the past. This benefit is akin to a BitTorrent player finding a better partner player than it had before, when it was trying to only search for cooperative behavior. Furthermore, optimistic unchoking could help with users who have very little data. If the Facebook algorithm player immediately began calculating MSI scores based on very little engagement data, it might show the user a limited variety of posts and never foray into posts that a user might also engage with or even prefer.

BitTorrent's randomization is often, but not too frequent – a new random peer is found every 30 seconds or so, which is a long time when downloading a file. Similarly, Facebook's algorithm could pair the user with a post that has an expected value of c*MSI where c = .5 every 30 posts or so. This ensures enough of a presence to somewhat alter a user's News Feed, but also does not severely undermine Facebook's goal of maximizing user engagement.

Why Is Expansion Good?

Although Facebook's algorithms player would not use randomization too frequently – it may pair the user with a post that had a c*MSI score every thirty posts – these random posts may make a meaningful difference. A study looking at echo chambers on Facebook found that randomizing entire News Feeds would have a significant effect on the political leanings of users' News Feeds: "If individuals acquired information from random others, ~45% of the hard content that liberals would be exposed to would be cross-cutting, compared with 40% for conservatives" (Bashky). Currently, 24% of liberals' News Feeds are cross-cutting hard content while that number is 35% for conservatives (Bashky). While my proposal would not entirely randomize News Feeds, it would introduce a significant random element. It could be reasonably expected that such a random expansion element would meaningfully add new kinds of posts to users' News Feeds.

Why do different kinds of posts – whether political or personal – matter? For many of its users, Facebook is the dominant source of news – news that shapes their political beliefs (Levy.) Expanding the kinds of posts users see, then, may affect users' political or world beliefs. While one may argue that a simple post would not be enough to change somebody’s mind, there is evidence that even cursorily reading a post is a learning experience for users. One study looking at user engagement on Facebook found that “Facebook’s News Feed, with its short article previews, provides enough information for learning to occur. This in itself is an important and normatively positive finding: in a relatively new way of acquiring information, Facebook users are learning by merely scrolling through their News Feed” (Anspach et al). The collective effect of multiple random posts, then, may be meaningfully processed by a user.

Of course, it is difficult to conjecture just how much adding a randomization element to Facebook’s News Feed would meaningfully impact users. There are no studies that have tested randomization, because such a mechanism does not exist. A field experiment, however, has come close. It classified Facebook users by their political leanings and encouraged them to like pages of opposing political viewpoints. Using a Chrome Extension, the study monitored which news sites and posts the users engaged with. Crucially, it found that randomizing content changes which news sites users visited and that “exposure to counter-attitudinal news decreases negative attitudes toward the opposing political party” (Levy). Of course, this study is different from my proposal because it just looked at exposure to opposite viewpoints. My proposal is unique because it is not just pushing material that is diametrically opposed towards users; rather, it is pushing material that may be orthogonal or slightly adjacent. The benefits to exposure, then, may be amplified: politically active users would be more likely to be receptive towards such material rather than polar-opposite messaging.

To understand what a diversity of information may look like, we could look to the Fairness Doctrine. The Fairness Doctrine was originally passed in 1949, mandating that radio and television programs had to display multiple sides of an argument. The logical reasoning was that radio and television programs operated on a limited amount of possible airwaves, thereby limiting the amount of possible market competition between programs (Fletcher). This precedent, however, did not apply to newspapers. In *Miami Herald Publishing Co. v. Tornillo,* the Supreme Court ruled that newspapers did not have to adhere to the Fairness Doctrine because there could be more market competition. And in 1987, under the Reagan administration, the policy was overturned. While there are few quantitative studies on the efficacy of the Fairness Doctrine, there are qualitative descriptions of its effect on public debate. By requiring different perspectives, “the Doctrine also protected the public rights of audiences to diverse information… Privileging public access to a rich marketplace of ideas over broadcasters’ rights was significant

— rarely have positive freedoms been so clearly articulated in U.S. legal and policy discourse" (Pickard). Creating a marketplace of ideas is a crucial solution to another problem on Facebook: content bubbles. Many users on Facebook may not be aware that they are in a content bubble and therefore are only consuming one side of the argument. A study looking at a sample of Facebook users in Germany found that passive users were more likely to be unaware that content bubbles existed; active users were more likely to be aware and desired for prevention strategies (Plettenberg). This indicates that even those who are aware of such bubbles are not sure what exists outside of them or how to access such content. In both cases, randomization performs an important function: making people aware of the diversity of arguments.

How does randomization tie into content moderation? First, groups will be less inclined to make their posts radical when anyone will be likely to see it rather than a small intended audience. Second, as people are exposed to more diverse content, they may be less likely to engage with radical posts. Political organizations and even private citizens may find that popularity is no longer a game of extremes. A less polarized public may reduce the incentive to post extremist content in the first place.

Randomization may also help Facebook's popularity (and therefore profitability) in the long term. Expanding information may be useful for Facebook for several reasons. First, creating a platform that is open to everyone, rather than just political extremists, is better for Facebook's long-term revenue. Currently, Facebook has prioritized a short-term model, valuing advertising revenue by increasing inflammatory content. As mentioned earlier, Mark Zuckerberg remains unwilling to sacrifice polarizing content for the sake of losing engagement. However, such a strategy will not fare in the long term. Public backlash has already begun to emerge: news outlets like the Wall Street Journal are publishing exposés such as The Facebook Files; Congressmembers from both parties lament the role of Facebook in rising extremism and wish to regulate it; users have even reported feeling like the platform has hurt the quality of their lives. Creating a platform where users are less polarized will serve Facebook's durability in the long term.

Conclusion:

As stated earlier, content moderation is typically conceived of as reducing what a user sees, not expanding what they see. However, expansion is a crucial part of content moderation, because effective content moderation addresses the root issues of why problematic content is created and popularized in the first place. Facebook's current approach to content moderation – taking down posts and regulating algorithms – will never solve the problem while its main algorithms continue to promote posts that contain the maximum MSI values. That is to say, Facebook's short-term ad-revenue business model is what makes the very enterprise of content moderation impossible. Adding random posts into a user's News Feed may help reduce negative sentiment which makes extreme posts popular and profitable in the first; as a result, it may begin to meaningfully address content moderation issues.

References

Anspach et al. *A little bit of knowledge: Facebook's News Feed and self-perceptions of knowledge.* Published in Research and Politics, January-March 2019.

Bashky et al. *Exposure to ideologically diverse news and opinion on Facebook.* Published in Science Journal.

Douek, Evelyn. *FACEBOOK'S "OVERSIGHT BOARD:" MOVE FAST WITH STABLE INFRASTRUCTURE AND HUMILITY.* Published in North Carolina Journal of Law & Technology.

Fletcher, Dan. *A brief history of The Fairness Doctrine.* TIME Magazine. http://content.time.com/time/nation/article/0,8599,1880786,00.html

Hagey, Keach. *Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead.* Published in The Wall Street Journal. https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215

Johnsen et al. *Peer-to-peer networking with BitTorrent.* Published by NTU, 2005.

Levy, R. *Social Media, News Consumption, and Polarization: Evidence from a Field Experiment*. American Economic Review. Published 2021.

NYT. *Facebook Has 50 Minutes of Your Time Each Day. It Wants More.* Published in the New York Times. https://www.nytimes.com/2016/05/06/business/facebook-bends-the-rules-of-audience-engagement-to-its-advantage.html

Pickard, Victor. *The Fairness Doctrine won't solve our problems — but it can foster needed debate.* Published in the Washington Post. https://www.washingtonpost.com/outlook/2021/02/04/fairness-doctrine-wont-solve-our-problems-it-can-foster-needed-debate/

Plettenberg N. et al. (2020) *User Behavior and Awareness of Filter Bubbles in Social Media.* In: Duffy V. (eds) Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Human Communication, Organization and Work. HCII 2020. Lecture Notes in Computer Science, vol 12199. Springer, Cham. https://doi.org/10.1007/978-3-030-49907-5_6

Savolainen, Petri. *Summary of the BitTorrent protocol.* (Emphasis mine.)

University of Iowa. *The BitTorrent Protocol.*