



Hateful By Design

Symptoms of and Solutions to Internet Radicalization

Cole Donat

Introduction

Anger is self-perpetuating at scale. Acts of anger inspire like outbursts, compounding toward a tipping point of societal violence. This paper explores the structure of social media, the internet as a whole, and the radicalizing effect upon its users. It closes by offering solutions to the systemic problems behind online radicalization and extremism.

Background

The ability of social media to radicalize users is rooted in the legal history of internet regulation. How the internet is regulated has shaped how we interact with social media as a whole, and how it can take such dangerous forms in the wrong environments.

Social Media is itself a novel phenomenon. It has existed for far less time than traditional media communications, such as cable TV or AM radio. As such, it exists in a state of near unregulation. This can be said for many fruits of the internet, which itself exists beyond the reaches of regulatory procedure. This unregulation is partially due to a lack of trying, but also reflects a legislative inability to understand the online world as an extension of the physical one.

This lack of direct regulation means the responsibility for content falls naturally to the platforms which host it, known as internet *intermediaries*. An intermediary can take many forms: a message board, a webpage, a video game, or a social media site. However, the difficulty of internet regulation begins with how legislation excuses intermediaries from the content they propagate: Section 230.

The year is 1996. The internet is still in its infancy, to a point almost unimaginable in modern terms. Daily online users have doubled from the prior year, 1995, to 20% of the

population¹. MySpace will not be invented for another 7 years². *Seinfeld* is still filming. The Democratic Clinton administration has just won reelection and is seeking a blockbuster bill to show it can make headway in the Republican 104th Congress.

Enter the Telecommunications Act of 1996; a wide-sweeping rewrite to the Communications Act of 1934. The Telecommunications Act was intended to reduced regulation to allow new communications technologies room to innovate and new competitors to enter the field³. Only a small section of the Telecommunications Act applied to the internet directly, appended as an afterthought later in the year under the name “the Communications Decency Act” (CDA). Even then, the CDA was not drafted with a focus on internet communication. Its purpose was to allow federal censorship of “patently offensive” content accessible to minors⁴. Within this post-script to appease anti-pornography lobbies came the cornerstone of all internet content regulation: Section 230.

Section 230 states:

“No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.”⁵

¹Auxier, B., & Anderson, M. (2021). Social media use in 2021. Pew Research Center, 1-4

² Britannica, The Editors of Encyclopaedia. "Myspace". Encyclopedia Britannica, 23 Feb. 2022, <https://www.britannica.com/topic/Myspace>.

³ Strickland, Ruth Ann. “Telecommunications Act of 1996.” Telecommunications Act of 1996, Middle Tennessee State University, <https://mtsu.edu>.

⁴ 47 U.S.C. Title IV

⁵ 47 U.S.C. § 230

In effect, this means that an open platform on the internet is not itself liable for content that users upload onto it. This was intended to reduce the cost of moderation faced by “internet computer service providers.” For example, if a user were to send out death threats via Match.com (notably founded 1995) to other users, Match.com would be protected from liability under Section 230 as they are not the publisher of information, merely a conduit (while the user responsible for the threats is still liable for their actions, law enforcement was, and is, unequipped to prosecute them online). As such, Match would not need to hire moderators to review and remove content, and could instead reinvest in server farms, software engineers, and digital marketing, thus expanding the nascent internet.

At the time, this certainly seemed to be a sensible decision. Many sites experienced too much data traffic to meaningfully moderate what users were uploading. The nature of user-generated content was also starkly different from previous media. Where a traditional broadcast such as television or radio could control all elements of its message, internet intermediaries were at the mercy of a tide of users.

However, online exchange was not entirely abandoned to anarchy. There was a substantial caveat where intellectual property found exclusion in the new bill: the Digital Millennium Copyright Act (DMCA).

The DMCA was passed in 1998, two years after the Communications Decency Act. This was another broad bill to implement new standards for copyright enforcement in the increasingly online age. Of specific focus to intermediary liability was the DMCA’s Online Copyright Infringement Liability Limitation Act, within it, Section 512⁶. Section 512 was designed to

⁶ 17 U.S. Code § 512

protect copyrighted information from being freely shared on the internet⁷. Until this point, online content moderation was considered unenforceable. There was an unmanageable amount of data shared under the unimpeachable anonymity of users. This led to a new way of thinking about online content moderation: the limited liability model.

The limited liability model, otherwise known as the “safe harbor” approach, was a way of punishing structural intellectual property mismanagement rather than individual instances⁸. If an intermediary had the necessary systems in place to expediently remove a copyright infringement once alerted, it would be protected from damages stemming from said infringement. For example, if a music publisher came to realize that a song of theirs was being used without permission on a social network, they could contact the administrators by filing a DMCA copyright strike. If the website administrator acts “expeditiously” to remove the content, they are immune from paying monetary relief to the music publisher.

This came to be known as the “report and remove” system. As the limited liability model became adapted internationally, a report and remove system would be a prerequisite for immunity in all cases⁹. Even though it could only be used reactively, it saw meaningful efficacy

⁷ “Digital Millennium Copyright Act.” Legal Information Institute, Cornell University, https://www.law.cornell.edu/wex/digital_millennium_copyright_act.

⁸ It is worth noting at this point that “intermediary” was substantially less defined. Much of Section 512 serves to protect ISPs such as Verizon Wireless, or server owners, from the content shared through their infrastructure. From this point on, intermediaries would come to mean website hosts such Facebook or Google.

⁹ Stoltz, Mitch. “Copyright's Safe Harbors Preserve What We Love about the Internet.” Electronic Frontier Foundation, 18 Jan. 2019, [Electronic Freedom Frontier](#).

in reducing the number of copyright infringements shared online. This showed that the responsibility of online moderation could be delegated in part to users, making the impossible task of moderating online content slightly more realistic.

The implementation of Section 230 and Section 512 created substantially different enforcement patterns for personal users and intellectual proprietors. On one hand, Section 512 delineated a clear course of enforceable action to remove copyright infringements, and leveraged the specter of civil liability over intermediaries to keep them responsive. On the other, Section 230 eliminated the responsibility of intermediaries to enforce other illegal forms of speech such as explicit threats and premeditated terrorist activity without judicial intervention. In conjunction, Section 512 and 230 would tilt the profit incentive of the burgeoning social media firms towards property interests and away from individual user safety.

Major Players in Social Media

The passage of 230 allowed the sector of Social Networking to bloom into a multibillion-dollar industry in the last decade. In 2010, 70% of the market share among social media firms was attributed to Facebook. Today, the top 78% of market share is split by Facebook (36%), Youtube (27.01%), Twitter (6.82%), Reddit (5.10%), and Instagram (2.47%)¹⁰. Also note that Facebook and Instagram are owned by the same parent corporation: Meta.

However, market share is not synonymous with use when it comes to networking. In the last year, Facebook has remained dominant in terms of visits, making up ~74% of visits out of all social networking sites. For reference, the next most visited site: Twitter, made up only 8% of

¹⁰ Karl. "Top 10 Social Networking Sites by Market Share Statistics [2022]." Dreamgrow, 17 Jan. 2022, <https://www.dreamgrow.com/>.

social networking visits¹¹. The staggering majority of visits to Facebook account for a massive base of daily users. Facebook reported that it hosted nearly 2 billion active users each day during the first fiscal quarter of 2022¹². As such, Facebook is nearly synonymous with the Social Media industry.

Industry Standards and Vocabulary

Radicalization over social media is shaped by the profit incentives of the intermediaries. An intermediary with an economic incentive toward maintaining a large, engaged user base will moderate differently than one incentivized to remove harmful content expeditiously. This leads directly into the business model of a social media firm.

Of the largest social media hosts, none charge subscriptions or access fees¹³. This indicates individual users are not customers. To this point, Facebook generates no revenue from users directly, instead relying on advertisers to purchase ads on their platform¹⁴. In the last fiscal period, Meta (Facebook's parent corporation) reported that 98% of their revenue came from advertisers buying advertising space on Facebook and Instagram. In this near universally used

¹¹ "Desktop, Mobile; Tablet Social Media Stats United States of America." StatCounter Global Stats, <https://gs.statcounter.com/>.

¹² Statista Research Department, and Apr 28. "Facebook: Daily Active Users Worldwide 2021." Statista, 28 Apr. 2022, <https://www.statista.com/>.

¹³ Falch, M., Henten, A., Tadayoni, R., & Windekilde, I. M. (2009). Business models in social networking.

¹⁴ "Meta Reports Fourth Quarter and Full Year 2021 Results." Meta, 2 Feb. 2022, <https://investor.fb.com/>.

model, users are not customers; they are products. User attention is sold to the Social Network's true customers—advertisers. All social network firms' behavior is ultimately shaped by their ability to attract advertisers, and in doing so, remain solvent.

To create more valuable advertisement space, users must be engaged with the content they are seeing. Advertisers are willing to pay a higher premium if metrics reflect that users interact with a large number of posts.

Of course, this creates a challenge for social networking firms. None of these firms create content themselves, and as such are unequipped to produce engaging content outright. Moreover, the vast majority of users consume more content than they produce. This means that if users can only view content from pre-approved accounts (such as friends/followers), they will quickly run out of content to view and in turn lose interest in the platform.

To combat the stagnation and exhaustion of platform content, networking firms must widen the pool from which users draw content. This typically manifests itself in a “recommended content” tab or page, or simply direct implementation into a feed. In recent usage, many of these features are referred to as a “For You Page”.

Of course, providing each user the same recommended content would represent a missed opportunity. Advertisers want to see more than passive scrolling; they want to see engagement. Engagement is measured in how often a user likes, reacts, comments, shares, or otherwise actively responds to a post. This shows attention, which is a more valuable advertising metric than conventional impressions or reach. This business model, geared toward generating engaged users for advertisers, has come to be known as “Engagement Marketing”¹⁵.

¹⁵ Harmeling, Colleen M., et al. "Toward a theory of customer engagement marketing." *Journal of the Academy of marketing science* 45.3 (2017): 312-335.

The rise of social media use has given Engagement Marketing incredible tools to grow. For context, Engagement Marketing first took hold in traditional mass media using representative samples, such as Nielsen IAG ratings¹⁶. As these were effectively large-scale customer surveys, Engagement Marketing was expensive to measure, relegating it largely to national programming by well-endowed firms¹⁷. The contrast between Nielsen ratings and modern social media enabled metrics is sharp.

Social media metrics are not samples. They are the entire picture. Data is kept on each user to show that they are engaged and to recommend content that will keep them so. There is no margin of error, nor lengthy and expensive searches. User attention metrics are infallibly updated in real-time, making the price of engagement tradable on the open market.

To keep users attentive, recommended content is tailored to each individual. This is the true lifeblood of a social networking firm. The more users that can be kept attentive through recommended content, the more advertising revenue can be generated. As such, the essence of social media is the ability to recommend content to users that they will engage with.

The process by which content is recommended is therefore the primary concern of a social media firm. Simply keeping in touch with established friends can be done over text, email, or other conventional methods. Firms must offer users new, engaging content to users. Across the

¹⁶Trunfio, M., & Rossi, S. (2021). Conceptualizing and measuring social media engagement: A systematic literature review. *Italian Journal of Marketing*, 2021(3), 267-292.

¹⁷Adgate, Brad. "For Nielsen Ratings Complaints and Potential Competitors Is Nothing New." *Forbes*, Forbes Magazine, 1 Sept. 2021, <https://www.forbes.com/>.

board, this is done through the use of proprietary data analysis, colloquially known as “the algorithm”¹⁸.

Algorithmically recommended content takes as many factors as possible into consideration when aggregating content to appear on a feed. Many users may see content because it shares a keyword with a previously liked post, or because their friends had interacted with it. Some users may even see posts because they are tagged in a nearby area, or related to a recently used search term. Much of the competition in the social networking industry relies on finding engaging posts for users. As such, these recommendation algorithms must use all data available to them, from account names to online shopping history¹⁹.

Advertiser focus, engagement marketing, and content recommendations have given way to a new brand of online market: the “Attention Economy”. In it, content is administered to users as feed is to livestock. In turn, their attention is sold for the opportunity to place advertisements in the same feed. Even though the Attention Economy eliminates the burden of cost on users, it creates issues as it runs its natural course.

Mechanisms of Radicalization

By the second decade of the new century, a lack of responsible incentives and a rise of new business models allowed social networking uninhibited access to the public consciousness. The natural result of this is a rise in violence, caused by inflammatory online rhetoric. Multiple facets of the social networking environment work in tandem to push users toward radical action

¹⁸Will Oremus, Chris Alcantara. “How Facebook Shapes Your Feed.” The Washington Post, WP Company, 26 Oct. 2021, <https://www.washingtonpost.com/>.

¹⁹ “Privacy Matters: Facebook Shops.” Meta, 21 June 2021, <https://about.fb.com/>.

at higher rates than other media. This section explores the factors that create an angry, violent user.

Early Symptoms of Extremism

In June of 2015, self-radicalized white supremacist Dylan Roof entered the Emanuel African Episcopal Church and murdered nine African American worshippers, all strangers to him, in a racially motivated attack. After being captured alive, Roof would go on to name the Council of Conservative Citizens (CCC) as the impetus for the attack. The Council of Conservative Citizens is a hate group monitored by the Southern Poverty Law Center in cooperation with federal law enforcement. Dylann Roof, however, was unknown to both institutions and the CCC, as he was not a member. Roof was willing to gun down strangers, fully aware of the death penalty he would later receive, for a group he had never personally interacted with. This was due to a new, but increasingly frequent phenomenon: self-radicalization²⁰.

Self-radicalization is a process in which individuals adopt violent ideologies passively, meaning without the influence of a recruiter or targeted action on the part of a hate group. Roof

²⁰ ALFANO, M., CARTER, J., & CHEONG, M. (2018). Technological Seduction and Self-Radicalization. *Journal of the American Philosophical Association*, 4(3), 298-322.

stated that his descent into radicalization began with a Google search of “black on white crime”²¹.

In 2018, Alek Minassian killed 10 people in a car attack in Toronto. Before doing so he uploaded a Facebook post implicating the social networking site 4Chan, and a subreddit known as “r/Incel”. Again, Minassian was never personally approached by either group, nor inspired to specific action that would be legally considered premeditation to a terrorist attack²².

These two examples are symptoms of an increasing number of self-radicalized terrorist actions. Self-radicalized terrorists operate alone, causing their attacks to be referred to as “lone wolf” terror events. There have been 59 lone-wolf terror attacks in the U.S. between 1950 and 2010. There were 44 between 2010 and 2020²³. This shows an explosive increase in the frequency of attackers that operate beyond the standard hierarchy of a terror cell. In the case of Roof and Minassian, both were radicalized by an environment that can be accessed anywhere with an internet connection. The particular danger of self-radicalized terrorists over their conventional counterparts is their unpredictability and individuality.

In effect, this means that similar attacks can happen at any time, in any place. Moreover, other users who may have contributed to an atmosphere of radicalization are difficult to hold

²¹ Berman, Mark. “Prosecutors Say Dylann Roof 'Self-Radicalized' Online, Wrote Another Manifesto in Jail.” The Washington Post, WP Company, 26 Oct. 2021, <https://www.washingtonpost.com/>.

²² Dickson, EJ. “How the Toronto Van Attack Suspect Was Radicalized by Incels.” Rolling Stone, 27 Sept. 2019, <https://www.rollingstone.com/>.

²³ Worth, Katie. “Lone Wolf Attacks Are Becoming More Common -- and More Deadly.” PBS, Public Broadcasting Service, <https://www.pbs.org/>.

accountable. Putting aside the anonymity of internet use, apart from the attacker themselves, no illegal action was taken. It is not illegal to implant violent ideas in the minds of others. To be prosecuted, one must be an active element of the exact planning or execution.

This structure of terrorism through radicalization, as opposed to recruitment, has allowed terror groups to operate online with near impunity. In a 2016 report, the Southern Policy Law Center noted “[w]hite supremacists are increasingly opting to operate mainly online, where the danger of public exposure and embarrassment is far lower, where younger people tend to gather, and where it requires virtually no effort or cost to join in the conversation.”²⁴

However, hate groups responsible for radicalization are not untouchable. While criminal charges for accessory to radicalization are ineffective, intermediaries have the right to regulate their own platforms as they see fit. This right is also baked into Section 230 as a subparagraph, which reads as follows:

“No provider or user of an interactive computer service shall be held liable on account of any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected”

The terms of service of all major U.S. social networking firms include a “community guidelines” section that prohibits the use of hate speech²⁵. Yet, online hate speech has shown

²⁴ Mark Potok. “The Year in Hate and Extremism.” Southern Poverty Law Center, 1 Jan. 1970, <https://www.splcenter.org/>.

²⁵ “Hate Speech.” Transparency Center, Meta, <https://transparency.fb.com/policies/>.

steady annual increase despite being explicitly banned²⁶. Social media in particular has fostered a number of groups that have led to violent terror attacks, such as UniteTheRight and QAnon²⁷. It is estimated that roughly half of white supremacist organizations maintain a Facebook presence²⁸.

Facebook does have automatic content moderation in place, and in the last decade Meta has implemented site-wide changes to combat the spread of hate speech. However, not all hate groups traffic in detectable hate speech. Circumventing content filters is a common occurrence on all media sites, and many can be evaded as easily as repl@cing a letter in a word. Even in the event of successful detection and removal from the site, members of hate groups can simply create new accounts and return without missing a beat. With this in mind, it is clear radicalization must be stopped as quickly as it can take hold. It is much easier to halt the progress of a user's radicalization than to de-radicalize them from an extreme viewpoint.

Regrettably, social media firms are not only complicit in the radicalization pipeline, they are an active participant in it.

Incentives to Inspire Anger

²⁶ “Report: Online Hate Increasing against Minorities, Says Expert.” OHCHR, United Nations, 23 Mar. 2021, <https://www.ohchr.org/>.

²⁷ “Hateful and Conspiratorial Groups on Facebook.” Anti-Defamation League, 3 Aug. 2020, <https://www.adl.org/>.

²⁸ “White Supremacist Groups Are Thriving on Facebook.” Tech Transparency Project, 29 June 2020, <https://www.techtransparencyproject.org/>.

Engagement, the golden metric for all social media content, is difficult to measure. Is a like more engagement than a share? What about a comment? What if a user simply stares at the post longer than others? In 2015 Facebook was experiencing a crisis of statistics. Its content generation algorithm was being exploited by corporate entities. Many advertisers found that highly-liked, shared, and commented posts were more likely to go viral. Rather than pay Facebook directly for advertising, these accounts could simply buy user interaction through fake accounts and reach far more users for far less money. As a result, user engagement took a dive as fewer relevant posts came across feeds, displaced by quasi-advertisements.

To combat this, Facebook implemented a new metric—the “meaningful social interaction” (MSI). This was an addition to the existing recommendation algorithm in which posts from other human users were given priority in the recommendation feed. Moreover, posts that generated “conversation” were granted the highest standard. Other sites quickly followed suit, and before long the meaningful social interaction became an industry standard.

However, user-to-user interactions take very different forms than that of corporate ones. While this reduced the amount of say, sassy Wendy’s tweets that appeared in a feed, it replaced them with viral posts from strangers. This meant that users were increasingly interacting with each other, but were being measured with the same engagement metrics meant for advertisers. The unfortunate consequence of this was that posts causing controversy were given the gold standard of a post that “generates conversation”²⁹.

²⁹ Hagey, Keach, and Jeff Horwitz. “Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead.” The Wall Street Journal, Dow Jones & Company, 15 Sept. 2021, <https://www.wsj.com/articles/>.

As a general rule, the same user commenting twice on a single forum is not a sign of user enjoyment. It is far more likely that a user is taking part in an argument. Someone arguing about, say, the efficacy of vaccines, can engage with a single post multiple times, sometimes for hours on end. Of course, Facebook is not concerned about user enjoyment over engagement.

This is wonderful for attracting advertisers. In the fiscal quarter following the implementation of the meaningful social interaction, Facebook's ad revenue grew by 57%, the largest increase the company has ever seen³⁰. The Attention Economy had found its new favorite engagement metric in controversy-steeped "meaningful" social interaction.

The meaningful social interaction would go on to shape the content recommendation algorithm itself. The algorithm must be constantly updated to maintain user growth, and to do so requires testing. This testing, using small pockets of users as subjects, would need a metric by which to measure success. The choice, of course, was the MSI, being the new draw for advertising revenue.

As engineers dutifully worked to enhance their site, they did so with the intent to maximize MSI impressions as much as possible. The algorithm was tweaked so user posts were more likely to be commented on, reposted, or shared. In doing so, Facebook became better at recommending controversial, anger-inspiring content.

³⁰ "Meta Reports Fourth Quarter and Full Year 2021 Results." Meta, 2 Feb. 2022, <https://investor.fb.com/>.

Anger as Engagement

In 2018, Facebook implemented a “reaction” feature to posts to receive more specific user feedback, and recommend more personalized content. The reactions included the standard “like,” a “love,” as well as “haha,” “wow,” “sad,” and critically, “angry.” Feedback from these reactions would be a driving force for determining if a post was recommended to other users.

In a leak of internal memos two years later, it would be revealed that Facebook weighted the “anger” reaction five times more heavily than the next highest³¹. This meant posts which users responded angrily to were far more likely to be recommended to other users’ feeds. This was not selected randomly. Test data showed anger-inducing content received the most engagement³². Social media firms are under no obligation to leave their users feeling better. In fact, if anger can be used to spark engagement, they are incentivized to leave users as angry as possible.

It is important to note that the remainder of the content recommendation algorithm is working in full effect alongside the increased anger reaction. This means controversial content will still be personalized by previous factors, such as location and mutual connections. Even alone, the societal damage done by the emphasis on anger would be disruptive. However,

³¹ Merrill, Jeremy B., and Will Oremus. “Five Points for Anger, One for a ‘like’: How Facebook’s Formula Fostered Rage and Misinformation.” The Washington Post, WP Company, 26 Oct. 2021, <https://www.washingtonpost.com/>.

³² Morris, Loveday, et al. “Whistleblower Testimony and Facebook Papers Trigger Lawmaker Calls for Regulation.” The Washington Post, WP Company, 26 Oct. 2021, <https://www.washingtonpost.com/>.

engaging with personal controversy would prove to only be the first step on the path toward radicalization.

Rabbit Holes, and Why We Fall Down Them

While the flow of content must be engaging, it must also be constant. Users will not return to a platform if they are exposed to the same content repeatedly. This means that another important factor feeds into the content aggregation algorithm: post frequency.

As stated, the vast majority of users will consume much more content than they produce. Particularly in the case of highly specified topics, users require a large pool of content that is relevant to them in order to stay engaged with their feed. Each time a captivating topic is exhausted, engagement is lost. It is no easy feat to find what interests a user and keep them steadily supplied with new entertainment.

The primary way large networks can address this is to favor frequent uploaders. For example, let us say a user has a pattern of engaging with content related to the movie *Cars 2*.

While there may be very little content that is immediately relevant to *Cars 2*, this could show an interest in animation, or automobiles, or sports racing. In deciding what topics to recommend next, the algorithm will take content from the most frequently-uploaded account.³³ Keeping in the example, let us say that two accounts fit the criteria for recommendation: an automotive mechanic, who uploads videos of cars they service daily, or an animator, who uploads clips once or twice each month as their projects complete. All else being equal, the

³³ Peruta, A., & Shields, A. B. (2018). Marketing your university on social media: a content analysis of Facebook post types and formats. *Journal of Marketing for Higher Education*, 28(2), 175-191.

frequent, active, content creator (in this case, our mechanic) will be selected to enter the user's feed.

This is a well-documented phenomenon on all major platforms. Facebook, Twitter, Instagram, and even LinkedIn have a recommended posting frequency³⁴. This allows users to jump into interests completely until bored, growing tired of one topic and flipping to another like spitting out chewing gum, then unwrapping another piece.

A new interest can be addictive. When a user has interacted with two posts of the same nature, they are far more likely to pursue a third of the same nature³⁵. This is known as the “rabbit hole” effect. These periods of sustained interest can feel all-consuming to users. Some may find new hobbies, others an academic curiosity. However, when new interests are recommended based on controversy, these rabbit holes can lead to exceptionally dark places.

On October 28th, 2017, a post on 4chan from a user under the pseudonym “Q” claiming the national guard would be deployed to suppress the U.S. civilian populace would attract mild attention. Had that been the end of it, Q would have fizzled out as most other bad faith content creators do. However, Q posted the following day, and the day after as well. On November 3rd, Q

³⁴ “How Often to Post to Social Media in 2022.” Social Media Marketing & Management Dashboard, 4 Feb. 2022, <https://blog.hootsuite.com/>.

³⁵Woolley K, Sharif MA. Down a Rabbit Hole: How Prior Media Consumption Shapes Subsequent Media Consumption. *Journal of Marketing Research*. 2022;59(3):453-471.

began cross-posting to Facebook. They would go on the post over 400 times within the next two months,³⁶ with each post becoming more bigoted and conspiratorial.

The rabbit hole went extremely deep. Q themselves provided hours of primary content, and supporting posts for Q could have filled days of constant scrolling. One such word for Q is a “hyperposter”—a user who uploads a wildly disproportionate amount of content relative to others. By December of 2020, the “QAnon” movement would have swelled to millions, with conservative estimates putting the number of QAnon supporters at nearly one in seven Americans³⁷.

The growth of the QAnon movement was doubtless aided by Facebook and YouTube’s content recommendation algorithms. The Q movement provided lifetimes of content to viewers, and promoted intense engagement³⁸. This was capitalized on until QAnon-related content could

³⁶ Tian, Edward. “The Qanon Timeline: Four Years, 5,000 Drops and Countless Failed Prophecies.” Bellingcat, 7 Apr. 2021, <https://www.bellingcat.com/news/americas/2021/01/29/the-qanon-timeline/>.

³⁷ Nickelsberg, Robert. “Study Finds Nearly One-in-Five Americans Believe Qanon Conspiracy Theories.” NBCNews.com, NBCUniversal News Group, 27 May 2021, <https://www.nbcnews.com/politics/>.

³⁸ O’Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (2015). Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems. *Social Science Computer Review*, 33(4), 459–478. <https://doi.org/10.1177/0894439314555329>

displace all other posts in a user's feed³⁹. After all, if a user were to believe that their government was run by an evil cabal, they would likely show more interest in insurrection than *Cars 2*.

QAnon is likely the largest scale that a group has been radicalized in, but content recommendation based on frequency is a common path to radicalization. This is partially due to the nature of users who post frequently, but still act human enough to pass as a meaningful social interaction. These types of users are often propaganda engines.

A Brookings study of ISIS-affiliated Twitter accounts showed that the top 0.03% of ISIS-affiliated hyperposters were responsible for nearly 10% of all tweets made by ISIS accounts⁴⁰.

Extremists are well aware that frequent posting will lead to increased reach. Internal communications within terror groups show a conscious decision to create content en masse⁴¹. Recruitment for all extremist groups, from QAnon to ISIS, can be done almost automatically online. This can result in all forms of terror, from lone-wolf attacks to genocide.

Outcomes

The Capitol Insurrection

³⁹ CNN. "Revealed: How Facebook Promoted Qanon to a 'North Carolina Mom'." Facebook Watch, 18 Nov. 2021, <https://www.facebook.com/>.

⁴⁰ Berger, J. M., & Morgan, J. (2015). The ISIS Twitter Census: Defining and describing the population of ISIS supporters on Twitter.

⁴¹ Winter, C., Neumann, P., Meleagrou-Hitchens, A., Ranstorp, M., Vidino, L., & Fürst, J. (2020). Online extremism: research trends in internet activism, radicalization, and counter-strategies. *International Journal of Conflict and Violence (IJCIV)*, 14, 1-20.

The ramifications of online radicalization become more grievous as our lifestyles move more online. The engagement of anger and the convenience of rabbit holes affect all of us, whether they lead to personal radicalization or not.

To this end, online extremism threatens the stability of the nation. Misinformation will become an increasing part of national discourse and social action. The effects of this have already begun. Following the rise of QAnon and the loss of Donald Trump, hyperposters, such as Q, claimed that the results had been fabricated. Q's final upload, made shortly after the election had been called for Biden, simply read, "Are you willing to take back this country?"⁴².

Conspiratorial pages, as well as private groups on Facebook, Twitter, Reddit, and YouTube (as well as other platforms) organized to protest the results of the 2020 election under the name "Stop the Steal."

On January 6th, 2021, as Congress convened to certify the election results, the "Stop the Steal" crowd gathered outside the U.S. Capitol. Many carried firearms loaded with live ammunition. A few flew flags representative of the groups responsible for their attendance. Among the symbols flown by the mob were a "QAnon" banner and a "Kekistan" flag (a far-right symbol made specifically to denote an online presence)⁴³. By 3pm, the rioters would breach the inner sanctums of the Senate and House, with lawmakers only narrowly evacuated beforehand. In the wake of the attempted coup, five were left dead including a capitol police officer.

⁴²Tian, Edward. *QAnon Drop Dataset*, Bellingcat, February, 2022. [Dataset \(open links within at own risk, as they re-route to Q-forum\)](#)

⁴³ Staff. "Identifying Far-Right Symbols That Appeared at the U.S. Capitol Riot." The Washington Post, WP Company, 15 Jan. 2021, <https://www.washingtonpost.com/>.

The January 6th insurrection was integrally formed by online radicalization. In the subsequent months, the extent to which the fatal rally was organized using Facebook came to full light. It was found that misinformation shared on Facebook was a driving attendance factor for rioters^{44, 45}. Congressional investigations into the role of social media in the attack remain ongoing. This implicates all forms of social media, and shows that the need to stem internet radicalization is not abstract.

Rohingya Genocide

Internet radicalization is not limited to the domestic United States. Our inability to stem hateful online behaviors has serious international implications.

In 2015, Myanmar began a digitization campaign. The number of social media users exploded across the nation. Facebook had a much firmer grasp over the average Myanmar user than their U.S. counterpart. To many users in Myanmar, internet use was nearly synonymous with Facebook, which maintained a powerful monopoly in the region⁴⁶. In 2016, the meaningful social interaction metric took effect, compounding the issues of hate speech within the nation.

⁴⁴ Craig Silverman, Craig Timberg. “Facebook Hosted Surge of Misinformation and Insurrection Threats in Months Leading up to Jan. 6 Attack, Records Show.” ProPublica

⁴⁵ Timberg, Craig, et al. “Inside Facebook, Jan. 6 Violence Fueled Anger, Regret over Missed Warning Signs.” The Washington Post, WP Company, 29 Oct. 2021, <https://www.washingtonpost.com/>.

⁴⁶ Angshuman Choudhury for The. “How Facebook Is Complicit in Myanmar's Attacks on Minorities.” The Diplomat, 25 Aug. 2020, <https://thediplomat.com/>.

One of the most common targets of hate speech in Myanmar were the Rohingya⁴⁷, a Muslim minority group concentrated in the province of Rakhine. A UN fact-finding investigation found Facebook's hatemongering algorithm was put on full display in a monopoly setting.

“By creating an environment where extremists' discourse can thrive, human rights violations are legitimised, and incitement to discrimination and violence facilitated.”

(UNCHR)⁴⁸

Rohingya activists pled with social networking firms, chief among them Facebook, to stem the flow of hate speech toward them. However, Facebook lacked the personnel to adequately moderate Myanmar. At the time, only two content moderators spoke Burmese⁴⁹.

The frequency and severity of the anti-Rohingya rhetoric increased for the next year, until in August of 2017, military and vigilante groups entered Rakhine. The ensuing violence would take 43,000 Rohingya lives, and displace three-quarters of a million others. The United Nations would go on to consider this a “textbook case of ethnic cleansing.” Military plans for the genocide were posted to Facebook by higher government officials the day before they were executed. Facebook removed them nearly immediately.

⁴⁷ Stevenson, Alexandra. “Facebook Admits It Was Used to Incite Violence in Myanmar.” The New York Times, 6 Nov. 2018, <https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html>.

⁴⁸ Nations, United. A/HRC/39/64 “Myanmar: Un Fact-Finding Mission Releases Its Full Account of Massive Violations by Military in Rakhine, Kachin and Shan States.” OHCHR, 18 Sept. 2018, <https://www.ohchr.org/>.

⁴⁹ Mozur, Paul. “A Genocide Incited on Facebook, with Posts from Myanmar's Military.” The New York Times, 15 Oct. 2018, <https://www.nytimes.com/>.

Radicalization is not just an individual issue. How a society interacts online affects how they act in person. It is of paramount importance that online forums are managed in responsible ways, with responsible incentives.

Possible Solutions

Rewriting Section 230

The most straightforward way to address online radicalization is legislative. Section 230 must be substantially rewritten.

Treating social media firms as independent from their content makes less sense in the modern context than it did in 1996. Today, social media relies on actively promoting user-generated content, not simply allowing others to view it. Social media firms act almost identically to a publisher for select user-uploaded content by disseminating, propagating, and profiting from it. Firms are using this content for economic gain, and should have to accept some responsibility for it.

In fact, even in 1998 it was clear that to promote content, one had to assume some liability for its effects. According to Section 512, “[The Intermediary will be immune to copyright damages if] the service provider does not select the recipients of the material except as an automatic response to the request of another person.” There exists precedent for legislature limiting content recommendation, although only for intellectual property. It should therefore be possible to add a clause to Section 230 exposing a social network to civil, or even criminal, liability if their algorithm recommends content that inspires a terror attack.

This would create a limited liability model for radicalization that is similar to intellectual property. This would allow large networks to maintain a healthy profit margin without sacrificing victims' leverage against those networks. From here, the need to find a solution to radicalization is passed to social networking firms, who are the best equipped to find a lasting technological solution.

Redefining Metrics

Many of the systematic issues causing radicalization can be rooted in how we measure engagement. By finding a new metric for engagement, it may be possible to turn the incentive of social media toward unity, instead of division.

The meaningful social interaction is a flawed measurement that encourages controversy. Advertisers may not want to be associated with divisive content. As such, if the will of the market were to turn to a different metric, such as the “aww” reaction for example, it stands to reason that the role anger takes in attracting advertisers would drop precipitously.

Moreover, the general measure of engagement does not need to substantially change to reduce radicalization, data analysis simply has to become better at determining true interest from anger-based reactions. Language and image detection technologies have already made promising inroads at detecting everything from hate speech to misinformation. It is possible to separate what makes users angry and what keeps them interested.

These seem like small tweaks, but make no mistake, this would be a revolutionary pivot for social media. All empirical data for measuring success would be made effectively useless. This would mean revisiting customer profiles, reviewing business models, and retraining neural

networks. Redefining what success looks like in the social media industry would be a hardship, but one that must be undertaken.

Addressing Existing Social Issues

It can seem as if online radicalization occurs entirely on its own. But, to wit, no online extremist has acted because they have no issue with the tangible world. Racial tensions and political unrest existed long before the invention of Facebook, and if the servers were to power down today, there would still be strife tomorrow. Online radicalization is dangerous because it magnifies our social issues. It ensures that for every divisive event, there is a dark, angry rabbit hole for users to fall into.

If we were able to fully address the countless conflicts plaguing our human experience, there would not be the same anger online. Even in incomplete measures, the more empathy users at risk of radicalization can exercise, the less likely they are to fall into that rabbit hole, and the less likely they are to take others with them.

This is, by far, the most unrealistic solution to online radicalization. However, it is still one worth noting. The problems of the online world do not exist in a vacuum. They can be addressed without ever turning on a screen. It is a lofty ideal, but the only long-term solution we have.

Conclusion

Many people find themselves soldiers in an army without direction, fighting a culture war that never was. Our anger has been exploited for private profit. Our legislators and business leaders must work to regulate an industry that evolved beyond its regulation. There is no longer a

difference between content and statement, nor publisher and platform. To continue to treat either as distinct will cause worldwide violence.

In the name of advertisement, internet users, particularly on social media, have been incensed to action. And to keep them enraged, had their worldviews narrowed to the point of blindness. We are now products. We are not at the table; we are on the menu. Users will continue to be exploited for fiscal gain until we have leverage against social media giants.

There is some debate over whether social media firms are participatory or merely complicit in radicalization. For now, this is a moot point. Regardless of intent, social media has shaken nations and taken lives. The systems which we only just laid the groundwork for in the last two decades must be reworked.

Fixing our dilemma is not an impossible task, though. Many of our problems stem from outdated legislation, updating even a few paragraphs will revolutionize the entire internet. Social media firms are balanced on harmful metrics, shaking them will make firms rethink. Finally, we have the capability to come together as a nation. Divide causes divide. Unity inspires unity. While there is neither excuse nor apology acceptable from violent extremists, the pipeline to the bitter end can be closed.