Securing Veracity: Defenses Against the Pervasive Influence of Social Media

JACOB POLATTY

Advisor: John Savage
Reader: Timothy Edgar

Submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science with honors at Brown University

May 4, 2022
Abstract

Securing Veracity: Defenses Against the Pervasive Influence of Social Media

by Jacob Polatty

While the evolution of social media over the past two decades has undoubtedly helped pave the way towards a global state of total digital connectivity, the unchecked growth of these platforms has inadvertently brought about a broad range of detrimental effects for both individual users and the stability of society as a whole. The gravity of these dangers became evident in 2016 following the revelations of foreign interference in the U.S. presidential election, and the breadth of the repercussions of social media on political affairs and international security has grown increasingly apparent in the intervening years. Through an analysis of leaked information and compulsory transparency reports from major technology firms, the primary causes for these threats posed by social media can be attributed to a combination of negligent executive decisions and inadequate algorithmic systems that contribute to the circulation of harmful content. This thesis includes a thorough examination of these critical events and disclosures to construct an account of the most pressing deficiencies within modern social media platforms, which serves as a foundation upon which to scrutinize methods for possible improvement. Building upon this historical framework, the remainder of this paper synthesizes a set of technical suggestions for achieving better protections against malicious conduct and outlines an original approach for improving automated moderation by inverting the results of recommendation algorithms to function as a signal for the presence of negative content. Finally, this thesis concludes with a selection of policy-based strategies for compelling companies to establish heightened standards of enforcement, yielding a complete proposal for the comprehensive reform of social media.
Acknowledgements

To my thesis advisor, Dr. John Savage, thank you for your generosity in working with me this year and for all of your contributions. My fascination with multidisciplinary solutions to security and misinformation threats was sparked by your Cybersecurity and International Relations course my freshman year, and I cannot overstate how much of an influence your teachings have had on the trajectory of my education. I am incredibly grateful for your continual support and kindness over the past four years and throughout the thesis-writing process.

To my reader, Professor Timothy Edgar, thank you for all of your assistance in helping me refine my initial ideas for my thesis into a concrete argument. I am particularly thankful for getting the chance to serve as your Head TA for Computers, Freedom, and Privacy last fall, which was a foundational part of my undergraduate experience and helped strengthen my ability to analyze issues at the intersection of technology and legal policy.

To Dr. Ugur Cetintemel, thank you for all of your guidance throughout my entire time at Brown and for the opportunities to engage in research. Working under your supervision was an invaluable learning experience and fostered my knowledge of machine learning, which informed a significant amount of the technical analysis for my thesis.

To my parents, thank you for your unending support through all of my endeavors in college. I could never have accomplished any of this without your reassurance and encouragement and I truly can’t thank you enough.
# Contents

1 Introduction 1

2 Historical Overview of Social Media 5
   2.1 Early Social Networking Platforms ................................. 5
   2.2 The Advent of Targeted Online Advertising ...................... 7
   2.3 The Rise of Personal Data Collection and Modern Advertising .... 10

3 Modern Disinformation Campaigns on Social Media 15
   3.1 Origins of Digital Disinformation .................................. 15
   3.2 Foreign Interference in the 2016 U.S. Presidential Election .......... 17
   3.3 Public Health Misinformation and the COVID-19 Pandemic .......... 21

4 The Facebook Files: An Inside Look at Facebook’s Failures 26
   4.1 Background on the Leaks .............................................. 26
   4.2 XCheck and the Inequity of Site Rules .............................. 27
   4.3 Treatment of Political Content After 2016 .......................... 30
   4.4 Influence of Facebook on the January 6th Insurrection ............. 32
   4.5 Violence and Censorship in Developing Countries .................. 35
   4.6 Racial Bias and Hate Speech .......................... 38

5 Moderation and Recommendation Algorithms 42
   5.1 The Uses of Algorithms on Social Media Platforms ................. 42
   5.2 The Legal Standing of Moderation .................................. 44
   5.3 Algorithmic Moderation Techniques ................................. 46
   5.4 A False Sense of Security from Ineffective Moderation .......... 50
   5.5 Recommendation Technologies and Algorithmic Amplification ..... 52
   5.6 Pitfalls and Manipulation of Addictive Recommendation Algorithms ... 55

6 Technical Mitigations for Algorithmic Challenges 60
   6.1 Tactics for Improving Moderation and Recommendation Algorithms ... 60
   6.2 Building Upon Existing Mitigation Strategies ..................... 62
   6.3 Media Literacy and Authoritative Content Measures ................ 65
   6.4 Recommendation Scoring as an Inverted Indicator of Legitimacy .... 69

7 Conclusions and Policy Proposals 73
   7.1 Lacking Self-Regulation and the Necessity of Policy Reforms ....... 73
   7.2 Ineffective Attempts at Legislating Social Media ................... 75
   7.3 Limiting Revenue Incentives and Compelling Maximum Enforcement ... 78

Bibliography 82
Chapter 1

Introduction

“And that means that we need to sort of give you a little dopamine hit every once in a while, because someone liked or commented on a photo or a post or whatever. And that’s going to get you to contribute more content, and that’s going to get you more likes and comments. It’s a social-validation feedback loop, exactly the kind of thing that a hacker like myself would come up with, because you’re exploiting a vulnerability in human psychology. The inventors, creators — it’s me, it’s Mark [Zuckerberg], it’s Kevin Systrom on Instagram, it’s all of these people — understood this consciously. And we did it anyway.”

While this quote from former Facebook president Sean Parker in a 2017 interview with Axios may not appear noteworthy in the wake of Facebook’s many scandals in recent years, his admission that their site was inherently designed to manipulate and alter the behavior of its users actually provided a deep, revealing insight into the inner workings of social media platforms and their relationship with their user bases. The extent of the influence of social media platforms including Facebook, Twitter, and Instagram over all facets of modern life cannot be overstated, with recent statistics showing that over 72% of American adults actively use at least one social media platform. Outside of the United States these numbers are even more staggering, with these major platforms having a combined global user population of 4.62 billion as of January 2022, comprising

---


a vast majority of the world’s 4.95 billion internet-connected individuals. But with this ubiquity comes an unprecedented degree of power over the course of world affairs and all domains of public discourse, and the dangers of this influence have been documented extensively. The scope of the threat that these platforms pose to the stability of world governments and societies has been particularly highlighted over the past six years, including through the investigations into foreign disinformation campaigns during the 2016 U.S. presidential election, the ongoing spread of deceptive public health information during the COVID-19 pandemic, and the body of leaked documents provided by former Facebook employee Frances Haugen in her 2021 whistleblower complaints. Although these abuses may be viewed as merely some kind of status quo within a tech industry that frequently prioritizes profit over individual privacy and safety, it is important to consider the factors that led to social media platforms being designed in the manner described above by Sean Parker. This discussion begins in Chapter 2 with a historical overview of the major events and design choices within the nascent development of corporations such as Google and Facebook that shaped the structure of the modern internet, focusing on the emergence of the industry standard of treating user data as an end product and economic factors such as targeted advertising that pressured these corporations to develop increasingly invasive data collection techniques in order to stand out from their competitors.

With this background in place, I then proceed into a detailed examination of the current state of affairs on social media with regards to handling disinformation and other forms of harmful content such as hate speech, focusing on the documented consequences of these corporate decisions which have directly led to the destabilization of world governments and borne a tangible cost in human lives. This analysis is divided into two parts, with the first comprising a pair of case studies regarding the impacts of social media on the 2016 presidential election and the COVID-19 pandemic, while the second will center solely on Facebook and their harmful corporate practices and culture that

---

were exposed through a series of leaks throughout the latter half of 2021.

After identifying the root causes of the substantial impacts of social media on society and placing a spotlight on these companies’ negligence in attempting to mitigate these abuses, the remainder of the thesis will shift to presenting possible methods to directly resolve these issues or help mitigate the effects of the structural problems that are endemic to social media. Although all of these companies employ some number of personnel dedicated to content moderation, the sheer scale of these platforms and the volume of content necessary to review forces them to rely predominantly upon automated recommendation and moderation algorithms that dictate what content is displayed to users. While all of the major platforms largely treat the implementation details of their algorithms as critical company secrets, a combination of leaked details and inferences made from external analyses has provided a significant degree of insight into their inner workings and has also highlighted cases where companies have neglected to leverage the maximum potential of their technical capabilities towards protecting the users of their platform. In some cases the prevalence of misleading content has even increased in spite of algorithmic changes designed to eliminate or prevent it from being automatically recommended to users. However, this failure is primarily rooted in an underlying current of competitive pressure, and these companies cannot be trusted to actually implement their optimal safety measures because any attempt to do so would likely undercut their main revenue stream.

While calls for reform to social media have frequently filled news headlines and political statements in recent years, little concrete action has been taken to regulate or provide additional oversight over any of the major social media platforms and federal regulation seems necessary in order to provide external oversight and a framework for forcing these corporations to adhere to certain rules and transparency expectations. Although most of the major social media companies have expressed their own verbal commitments to combat disinformation and enact measures to limit the spread of deceptive material on their sites, these efforts have failed to effectively decrease the pervasiveness of disinformation on their platforms. As a result of the lack of successful reforms by these major firms,
any attempt to resolve this multi-faceted challenge must first consider the measures that have already been implemented and scrutinize the reasons for which many of these companies have limited the overall scope of their algorithmic interventions and avoided imposing more substantial changes to diminish the potency of disinformation. Due to the dearth of incentives for Facebook to impose internal oversight on their recommendation algorithms to suppress misleading or blatantly inaccurate content, it seems necessary for these issues to be brought into the legislative domain, and in the final chapter I consider and propose potential legislative measures for laying out definitive expectations for all social media platforms to adhere to as part of a national agenda for mitigating disinformation.
Chapter 2

Historical Overview of Social Media

2.1 Early Social Networking Platforms

Before considering the modern effects of social media and its extensive influence over society, it is important to first analyze the initial development stages of these platforms in order to provide a background on the decisions that shaped this domain. Through this process, it is possible to identify many of the key design choices or features that jumpstarted the social media industry and are still responsible for many of the abuses that are carried out on the modern implementations of these sites. Furthermore, this discussion highlights how these companies invested the vast majority of their energy and development work into maximizing rapid expansion and achieving the largest possible market share of users, sparking strong competition between many of the early players in this field. After experiencing an initial degree of success, many of the primitive services in this space were ultimately being usurped by the rise of more robust platforms such as Facebook and Twitter during the 2000s, and the dominance of these newer corporations has persisted to this day.

Although the majority of the discussion surrounding social networks has focused on the platforms that have arisen since the public release of the World Wide Web in 1993, it is important to note that the idea of using internet technologies for networked communication actually originated nearly two decades before.\(^5\) This development was primarily driven through the 1970s by college research teams working to create simple

---

Chapter 2. Historical Overview of Social Media

bulletin board systems and forums that would facilitate discussion between computer terminals, culminating in the introduction of USENET in 1979 by a team of graduate students at Duke University and University of North Carolina. USENET was structured as a decentralized collection of local “news servers” that communicated with users via the Unix-to-Unix Copy Protocol (UUCP) rather than a single cohesive network, and the system’s versatility and widespread adoption helped to drive the design of web-based messaging systems moving forward.

While USENET provided the foundation for the core discussion and information-sharing features of social media, particularly relating to user interactions over shared topics of interest, a complete implementation of a centralized worldwide social platform did not arrive until the late 1990s. Technically, a full-fledged social networking system can be defined as requiring the capability for users to “(1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system,” in addition to providing some form of messaging capabilities for users to interact with their connections. The first internet service that met all of these stipulations was the rudimentary network SixDegrees.com, which was only active from 1997 to 2000 but served as the initial standard for establishing the trends of public profile pages and Friends lists to maximize the potential for users to connect. SixDegrees.com was originally created as a means to demonstrate the relations between individuals online inspired by the theory of six degrees of separation, which claims that any pair of individuals in the world can be linked by a chain of at most six social connections.

While this notion was initially created as a mathematical thought experiment, it reflects a fundamental structure within the nature of human communication that can be uniquely

---

7Encyclopedia Britannica, USENET.
modeled through social networks.

In spite of the failure of the SixDegrees concept, the social media domain began to explode in the early 2000s, driven by huge increases in the size of the internet-connected population and an increased interest in using digital technology for social purposes. One particular feature of this developmental stage was the increased specialization displayed across many of the platform. This ranged from services aimed at businesses and professionals including LinkedIn and Ryze.com to platforms like Friendster designed to help users reconnect with past acquaintances or sites like Flickr and YouTube focused on a single area of content sharing such as photos or videos. \(^{11}\) This specificity was actually shared by Facebook in its early days, with the platform initially being restricted to just college students through its first two years of existence before its collegiate expansion and significant funding prompted the site to complete its public launch in 2006. \(^{12}\) Subsequent major platforms such as Twitter and Instagram progressed directly to a global release and provided support for a wide range of content types and discussion topics from their inception, leveraging the massive success of Facebook as a new blueprint for the creation of a prosperous social network.

2.2 The Advent of Targeted Online Advertising

While the historical perspective described in the previous section provided a general overview of the evolution of social media and its features within the early years of the industry, it is also critical to consider the process by which these platforms transformed into highly profitable ventures. The period of time from the late 1990s to the early 2000s that featured the rapid rise of social media also served as the foundational era for the establishment of the modern internet economy; many of the guiding principles and corporate strategies that are still used to this day are drawn from the culture and design choices of this primitive industry, particularly when considering the realm of

---


advertising and its use within digital technologies. Although some early platforms opted to adopt subscription-based or premium models for the use of their online systems, the predominant market trend in this era sought to adhere to the concept of the “free and open internet,” which entails that all online platforms should be publicly available at no direct cost to their users.\footnote{Free Press. \textit{Free & Open Internet}. URL: \url{https://www.freepress.net/issues/free-open-internet} (accessed 03/25/2022).} This standard can be considered to have played a major part in enabling the meteoric expansion of many early internet companies by allowing clients to try their services free of charge, though it also had the direct consequence of making the technology sector incredibly reliant upon advertising as the chief source of revenue.

The introduction of advertising to the internet occurred within a year of the launch of the World Wide Web in 1991 and rapidly grew in prevalence through the establishment of dedicated digital advertising firms seeking to capitalize on the financial promises of this novel domain. The concept of the web banner advertisement was first realized on October 27th, 1994 when a Wired magazine subsidiary sold a portion of their front page to AT&T in an experiment to see whether individuals could be convinced to follow a hyperlink in an ad to another website. Prompted by the banner text “Have you ever clicked your mouse right HERE? YOU WILL,” over 44\% of users accessing the Wired page clicked through from the advertisement and were sent to AT&T’s home page.\footnote{Karla Hesterberg. \textit{A Brief History of Online Advertising}. HubSpot. Nov. 28, 2021. URL: \url{https://blog.hubspot.com/marketing/history-of-online-advertising}.} While this ratio or “click-through rate (CTR)” was exceptionally high in comparison to the CTRs found in modern advertising campaigns, the results of this trial served to highlight the vast potential of digital advertising.\footnote{Maciej Zawadzinski and Michael Sweeney. \textit{The History of Digital Advertising Technology}. Clearcode. URL: \url{https://adtechbook.clearcode.cc/history-advertising-technology/} (accessed 03/25/2022).}

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{fig2.png}
\caption{Archived image of the first ever online banner advertisement purchased by AT&T on Hotwired.com.}
\end{figure}
Although the start of online advertising featured sites directly negotiating deals with corporate partners for the purchase of physical space on their pages, the increasing scale of the internet prompted the creation of independent advertising networks seeking to automate the process of online ad delivery. The focus on click-through rate as the primary metric for determining the success of an ad campaign pushed the industry to begin considering mechanisms for tailoring advertisements to the interests of their user base to achieve better engagement. This process initially began at the level of classifying entire sites based upon whether an advertiser’s “ideal consumers” were likely to visit them, with agencies such as WebConnect providing statistical analyses aimed at identifying locations where advertisements could be viewed by the largest populations of the company’s “target demographics.”

WebConnect was also responsible for the shift from providers purchasing static ad space for a specific time period to a recognition of the phenomenon of “banner fatigue” in which a user becomes accustomed to the appearance of a specific advertisement and begins to ignore it, with the CustomView technology shuffling ad displays in an effort to maintain user engagement.

The continued expansion of the internet and growing demands from advertising partners forced the industry to again redefine its metrics for successful campaigns following the WebConnect era through the introduction of more sophisticated return-on-investment (ROI) tools. These new industry standards were primarily driven by advertising firm DoubleClick and its innovative D.A.R.T (“Dynamic Advertising Reporting Targeting”) technology, which tracked a wide range of detailed data measures and provided advertisers with an unprecedented degree of insight into the performance of their ads. DoubleClick ultimately stood as the first complete implementation of a targeted advertising platform by granting its customers significant influence over the roll-out and propagation of their advertising campaigns, including an individuated degree of control over the presence of ads on each partner site based upon their results.
the success of DoubleClick led to a deliberate move away from CTR as the measure of advertising impact towards the institution of a new metric of “cost per thousand impressions (CPM),” recognizing the shifting standards in the digital advertising space and the general reduction in the rate of click-throughs across the internet.\textsuperscript{20}

A final area of development within this burgeoning domain occurred within the context of search engines, which were becoming increasingly necessary for navigating the expanding internet. These search engines provided a prime opportunity for new sources of revenue by selling top search result placements to advertising partners to leverage the inherent precedence of these positions and the increased likelihood that users will click on one of the top results. This concept was pioneered by the early platform GoTo.com in 1999 through the introduction of its “pay-for-placement” system, which was soon expanded into a more robust “pay-per-click” (PPC) approach that allowed for firms to bid on a fee that they would pay for each user that the search engine directed to their site.\textsuperscript{21} While this program proved to be highly lucrative and was largely responsible for the rise of search engines as one of the most powerful technology sectors in the early 2000s, it also solidified the industry standard of directly commodifying user interactions. In particular, these systems helped to establish the new trend of placing a price upon a service’s capability to influence user behavior and convince them to engage with their clients’ advertisements, effectively centering their entire financial model upon targeted advertising.

\section*{2.3 The Rise of Personal Data Collection and Modern Advertising}

Despite the initial economic promise of concepts such as pay-per-click advertising, the effectiveness of the first implementations of these advertising systems was noticeably limited due to the relative lack of detailed information regarding their users’ interests and tendencies to engage with different advertisements. While the use of PPC as the

\textsuperscript{20}OKO Ad Management, \textit{The History of Online Advertising}.

\textsuperscript{21}Hesterberg, \textit{A Brief History of Online Advertising}.
general framework for search prioritization pricing has persisted to this day, the structure of digital advertising was completely transformed in 2002 through Google’s launch of its new AdWords program that delivered results on a per-user basis rather than operating at a site-wide level.²² This seemingly minor change in focus was actually responsible for a titanic shift in the technology industry’s treatment of advertising services and the value of user data. Following Google’s introduction of targeted user-based advertising, companies immediately began moving to prioritize data collection as a means to maximize their advertising potential, effectively creating a digital arms race comprised of progressively invasive collection measures. The heightened competition during the early 2000s amongst both search engines and social media platforms was ultimately responsible for many of these surveillance techniques becoming standards within the digital sector and forcing internet platforms to engage in privacy abuses in order to maintain a viable business model.

Although the decision by Google to implement user-based targeted advertising may appear to have been part of a deliberate corporate strategy, it actually originated as a result of an accidental discovery of the extent of the data that was being stored incidentally by their Search platform. Google was founded in 1998 by Stanford Ph.D. students Sergey Brin and Larry Page as a data-driven search engine built upon the foundation of their groundbreaking PageRank algorithm, which provided a more nuanced and robust metric for prioritizing search results to challenge the more traditional approaches of existing platforms such as GoTo.com and Yahoo.²³ In spite of Google’s novel approach to the problem of processing queries and ranking webpages, their initial treatment of advertisements derived largely from prior applications and their original implementation of AdWords featured a similar pay-for-placement system to GoTo.com.²⁴ However, this status quo approach was fundamentally altered by the analysis of Search engineer Amit Patel, whose data mining work on search records revealed that the platform had been

²⁴ Hesterberg, *A Brief History of Online Advertising*. 
collecting information on “a wake of collateral data such as the number and pattern of search terms, how a query is phrased, spelling, punctuation, dwell times, click patterns, and location.” While other search engines and Google’s initial implementation relied solely upon the selection of keywords entered by the user as the sole basis for performing a query, Patel’s findings enabled the company to refine their Search program to incorporate these “broad sensor[s] of human behavior” into their algorithms and effectively begin a new paradigm of data usage in internet technologies.

While these substantive changes had the immediate effect of helping Google to surpass its main competitors in terms of search result quality and relevance, the scope of data collection also had the long term impact of providing a far wider range of attributes for informing a revised targeted advertising platform. Although Google’s initial storage of this data had been wholly incidental, Patel’s discovery sparked a new wave of intentional and systematic accumulation of data signals that “recognized the gold dust in the detritus of its interactions with its users and took the trouble to collect it up,” relying upon this “data exhaust” as the foundation of the company’s financial strategies. Bolstered by the success of incorporating these behavioral features into their core Search algorithms, Google turned its attention towards addressing its limited financial prospects by attempting to completely redefine the relationship between search engines and advertisers, with Larry Page asserting that their advertising partners “shouldn’t even get involved with choosing keywords—Google would choose them.”

The subsequent modifications to the AdWords system yielded the most complex advertising prioritization algorithm in internet history to date, relying upon a PPC framework but providing an advanced computation of the “price-per-click multiplied by Google’s estimate of the likelihood that someone will actually click on the ad.” The resulting implementation of the AdWords platform served not only to propel the rise of Google to the forefront of the search engine domain but also functioned as the single

---

largest transformation in the history of digital advertising, with the entire tech industry moving to implement their own forms of targeted user-based advertising in an effort to replicate Google’s striking success. The majority of these efforts became centered on the collection of “user profile information” to “increase the relevancy of ads served for some user request,” with these tech firms developing algorithms to engage in a range of surveillance and behavioral inference procedures in order to build out their descriptions of user identities.\textsuperscript{30}

However, while these data collection procedures were already invasive when used within the context of search engines, they became significantly more powerful and began to provide a far greater threat to digital privacy when they crossed over into the domain of social media through the adoption of targeted advertising systems by corporations including Facebook and Twitter. Rather than being forced to rely solely upon indirect behavioral signals, these platforms were able to integrate these auxiliary signals with the information directly provided by users through their public profile and the selection of pages that they follow and interact with to allow for even more precise targeting of advertisements towards specific individuals and groups.

These targeted approaches are actually found to yield far more lucrative advertising results, with Mark Zuckerberg personally admitting that Facebook’s strategy is “much less [about] increasing the volume of ads and much more about increasing the quality of the content and the quality of the targeting to get the right content to the right people.”\textsuperscript{31}

Due to the extreme specificity of targeting, these platforms are able to limit the overall prevalence of advertisements on their site in favor of hosting a smaller number of highly targeted ads that are incredibly pertinent to each individual user’s interests. Although this phenomenon could be construed as the greatest economic triumph of the digital age and is largely responsible for the continued widespread commitment to a free and open internet, these developments stand as an irreversible trend of valuing advertising profits above all other considerations and ultimately led to the permanent erosion of user

\textsuperscript{30}Zuboff, \textit{The Age of Surveillance Capitalism}, p. 78.

privacy on social media.
Chapter 3

Modern Disinformation Campaigns on Social Media

3.1 Origins of Digital Disinformation

Although the establishment of targeted advertising set the stage for many of the adverse effects of social media, the extent of the harmful influence that internet platforms could have on society did not become apparent until the late 2010s. The majority of the early discourse in this domain over sites such as Facebook and Twitter centered on their addictive nature and possible negative impacts on mental health through the facilitation of harassment and cyberbullying, and these issues stood as the primary challenges for these platforms throughout their first decade of its existence. This narrow focus was reflected within the contemporary research in this field, which typically attempted to discover the presence of any measurable links between the use of social networking services and heightened risks of negative psychological effects including stress and depression. Despite a vast body of colloquial evidence for the numerous ways in which social media usage can contribute to mental health problems, many studies failed to establish a statistically significant correlation between these trends and were unable to ascertain any definitive effects of social media on public health.

---


Through this process, the predominant discussions regarding methods to regulate or fix social media remained concentrated on addressing these micro-level impacts on the wellbeing of individual users, and comparatively little emphasis was placed on considering the latent macroscopic impacts that such platforms could have on the world at large. While examples began to arise of social media being leveraged to spread false reports or deceive others for personal gain, these were generally treated as isolated incidents rather than a portent of greater dangers. Early attempts to define this problem including the World Economic Forum’s 2013 report Digital Wildfires in a Hyperconnected World treated the idea of “massive digital misinformation” as largely theoretical rather than an immediate tangible risk, seemingly underestimating the magnitude of the threat these platforms could pose to the stability of society.  

Furthermore, the gravity of these ramifications on the legitimacy of public discourse continued to be downplayed even as research began to emerge regarding the ability for social media to impact the political sphere, with a foundational study by a team of data scientists from Northeastern in 2014 demonstrating a clear link between consumption of “alternative” digital news sources and beliefs in false or misleading political opinions during the 2013 general election in Italy. In particular, the researchers raised concerns about Facebook’s role in “bursting the diffusion of false beliefs when truthful and untruthful rumors coexist,” and highlighted how the general “antagonism” within radical groups towards mainstream media sources could result in the delegitimization of facts and prevent reconciliation by making “any kind of persuasion process, even if based on more solid information, very difficult.”

While the fears expressed by the Northeastern team received little attention at the time of their release, they proved to be a highly accurate prediction of the potential for social media to be used to influence political beliefs. This topic was ultimately thrust into the general consciousness two years later through the revelations of foreign

---

36 Mocanu et al., Collective attention in the age of (mis)information, p. 8.
interference in the 2016 U.S. Presidential Election, which dramatically redefined the public opinion of social media and opened a wave of inquiries into the mechanisms by which these influence operations were carried out. Much of the disinformation that occurred during the 2016 campaign had actually been anticipated by the World Economic Forum in their 2013 report, which warned of the perils of situations in which “information circulates within a bubble of likeminded people who may be resistant to attempts to correct it.” The WEF also warned of cases of “high tension, when false information or inaccurately presented imagery can cause damage before it is possible to propagate accurate information,” directly foreshadowing the events of the COVID-19 pandemic and the rampant misinformation over public health reports and the safety of vaccines.

The remainder of this chapter focuses on the 2016 election and COVID-19 pandemic as case studies of the extent to which social media can have a wide-reaching impact on the stability of nations and the wellbeing of the global populace, emphasizing the means by which these platforms have been employed to destabilize the public perception of truth.

3.2 Foreign Interference in the 2016 U.S. Presidential Election

The influence campaigns led by the Russian government to impact the results of the 2016 U.S. elections ultimately stood as the culmination of a decades-long effort to leverage internet technologies for the purposes of shaping political opinions, yet the nature and existence of these schemes did not begin to reach public awareness until the days following the election in November 2016. While the months leading up to the election had featured the disclosure and attribution of Russian involvement in the hacks against the Democratic National Committee (DNC) and Hillary Clinton’s campaign, contemporary statements from the U.S. intelligence community focused solely on these hacks and made no mention of the use of social media as an additional attack vector. The U.S. Intelligence

37 Howell, Global Risks 2013, p. 25.
38 Howell, Global Risks 2013, p. 25.
Community (USIC) and Department of Homeland Security (DHS) even identified how the release of documents obtained from these hacks matched Russian precedent and stated that “Russians have used similar tactics and techniques across Europe and Eurasia, for example, to influence public opinion there,” but failed to publicly address the potential of influence campaigns being executed through more public channels.\footnote{DHS Press Office. Joint Statement from the Department Of Homeland Security and Office of the Director of National Intelligence on Election Security. Oct. 7, 2016. URL: \url{https://www.dhs.gov/news/2016/10/07/joint-statement-department-homeland-security-and-office-director-national}.}

This lack of consideration for the role that social media could play in complementing the electoral interference through hacking stands as a clear oversight when considering Russia’s traditional methods of spreading propaganda. These campaigns largely derive from Cold War-era Soviet intelligence strategies and typically take the form of a “high-volume and multichannel” approach that is “rapid, continuous, and repetitive” and “lacks commitment to objective reality or consistency.”\footnote{Christopher Paul and Miriam Matthews. The Russian “Firehose of Falsehood” Propaganda Model: Why It Might Work and Options to Counter It. Rand Corporation. 2016. URL: \url{https://www.rand.org/pubs/perspectives/PE198.html}.} Under this so-called “firehose of falsehood” scheme, the attackers are able to effectively overwhelm all discourse within their targeted discussion and in the process have the effect of diluting the value of legitimate reports by generating confusion and obfuscating facts.\footnote{Paul and Matthews, The Russian “Firehose of Falsehood” Propaganda Model.}

Following in the footsteps of the Soviet “active measures” strategies for gaining an upper hand in international politics, this style of propaganda campaign was particularly well-suited to the format of social media through the use of coordinated networks of anonymous users.\footnote{Andrew Weisburd, Clint Watts, and JM Berger. Trolling for Trump: How Russia is Trying to Destroy Our Democracy. War on the Rocks. Nov. 6, 2016. URL: \url{https://warontherocks.com/2016/11/trolling-for-trump-how-russia-is-trying-to-destroy-our-democracy/}.} Russia had actually been detected to have been employing systems of troll and bot accounts since at least 2014 to defend their invasion of Crimea and support the Assad regime during the Syrian Civil War, yet the potential use of these capabilities to influence a foreign election was still given little attention.\footnote{Weisburd, Watts, and Berger, Trolling for Trump: How Russia is Trying to Destroy Our Democracy.} Moreover, Russia’s ideation of using digital technology for the purposes of influencing foreign affairs can actually be traced back to 1998, when Russian representatives to the United Nations expressed significant
3.2. Foreign Interference in the 2016 U.S. Presidential Election

concerns about the potential for the “creation of information weapons and the threat of information wars” being used with consequences “which may be comparable to that of weapons of mass destruction.”

Although the overall effort spanned years of planning and testing, the main Russian influence operation against the American election was relatively short-lived and only actively targeted voters during the primaries and general election. These efforts were spearheaded by the Internet Research Agency (IRA), a Kremlin-backed organization founded in 2013 with the general goal of interfering in American politics by “spread[ing] distrust towards the candidates and the political system in general.” However, while the Justice Department’s indictment of the IRA in 2018 specified how the group spent the years leading up to the election developing disinformation strategies, their workers were not ordered to begin disseminating electoral information until February 2016 when an internal memo specified that they should “use any opportunity to criticize Hillary and the rest (except Sanders and Trump—we support them).” The IRA relied upon a system of “thousands of botnets, teams of paid human ‘trolls,’ and networks of websites and social-media accounts” in order to accomplish this mission, flooding social media platforms with posts and links to fake stories aimed at promoting their favored candidates.

This approach proved to be exceedingly effective at naturally spreading disinformation to a massive population directly through recommendation systems such as Facebook’s News Feed, where a retroactive analysis discovered that propaganda posts from a network of just 6 IRA-run pages were shared over “340 million times,” and this

---


total represented just a “tiny sliver” of the overall scope of IRA operations.\footnote{Craig Timberg. “Russian propaganda may have been shared hundreds of millions of times, new research says”. In: The Washington Post (Oct. 5, 2017). URL: https://www.washingtonpost.com/news/the-switch/wp/2017/10/05/russian-propaganda-may-have-been-shared-hundreds-of-millions-of-times-new-research-says/.
}

In addition to the IRA’s efforts to circulate fallacious stories through direct sharing and algorithmic manipulation, the organization also relied upon online advertisements as a powerful tool to deliver directed political messaging outside of the traditional social media content flow. IRA operatives were recorded “concealing their Russian identities… through false personas” in order to “produce, purchase, and post advertisements on U.S. social media and other online sites expressly advocating for the election of then-candidate Trump or expressly opposing Clinton.”\footnote{United States of America v. Internet Research Agency, p. 19.
}

In particular, the IRA took advantage of the hyper-targeting capabilities of social media advertisements that were detailed in Section 2.3 in order to promote these positions directly to specific demographics and voter groups, focusing on critical areas in swing states in order to maximize their chances of shifting the results of the election.\footnote{Massimo Calabresi. “Inside Russia’s Social Media War on America”. In: Time (May 18, 2017). URL: https://time.com/4783932/inside-russia-social-media-war-america/.
}

Ultimately, during Congressional hearings on the interference campaigns Facebook reported that these advertisements purchased by the IRA had been viewed by “as many as 126 million users” in the 6 months preceding the general election, affecting almost 90% of the voting population in the 2016 election.\footnote{Eric Tucker and Mary Clare Jalonick. “Newly disclosed Facebook ads show Russia’s cyber intrusion”. In: AP News (Nov. 1, 2017). URL: https://apnews.com/article/north-america-donald-trump-ap-top-news-hillary-clinton-politics-343e66b94d614227a29ecc1f8ba8b848.
}

The sheer scale of this campaign reflects the fundamental danger inherent within the structure of targeted advertising technologies, with the ability to demarcate users into narrow homogeneous groups raising the potential for malicious actors to single out these populations and inundate them with manipulative content.

Despite the presence of immediate evidence concerning the widespread use of social media by Russian operatives to spread disinformation, the major social media corporations initially pushed back against the idea that they had played a role in influencing the results of the 2016 election. Mark Zuckerberg strongly dismissed the notion that “fake
news on Facebook...influenced the election in any way” as a “pretty crazy idea” in his first public appearance following the election, and further rejected the suggestion that his platform was contributing to political divisiveness through the formation of online echo chambers. While Facebook eventually retracted this position months later after internal research confirmed that “governments or other malicious non-state actors are using its social network to influence national elections,” their professed commitment to confront these abuses appeared as a capitulation to national backlash and Congressional pressure rather than a self-motivated drive to improve their platform.

Furthermore, empirical evaluations of the role of these platforms in the 2016 election generally concluded that “fake news most likely did have a substantial impact on the voting decisions of a strategically important set of voters,” directly corroborating the role of social media in facilitating electoral interference.

### 3.3 Public Health Misinformation and the COVID-19 Pandemic

While all the leading mainstream social networks made commitments in the wake of the 2016 election to enact changes to their platforms aimed at combating misinformation and inhibiting the spread of foreign influence campaigns, these reforms remained relatively untested over the following three years. However, the emergence of the COVID-19 pandemic resulted in a massive wave of misleading public health reports and intentional disinformation efforts from both domestic and foreign sources that directly tested the reliability of these new mitigation techniques. Many of these influence campaigns sought to take advantage of the widespread fears and confusion in the early days of the pandemic to evoke distrust in the medical community and instigate discord within the American populace during another election season. Although the responses from major platforms


such as Facebook and Twitter to this rising threat were far more rapid and decisive than
their post facto handling of the similar challenges in 2016, the scope and intensity of their
resolutions revealed some of the fundamental limitations within the structure of social
media for protecting against disinformation.\textsuperscript{56} The resulting commitment to provide
greater transparency in matters of public health further emphasized these deficiencies,
bringing to light areas of digital discourse that these firms are unable to adequately
regulate or control.

The emergence of misleading content relating to the novel coronavirus began almost
currently with the rise of public awareness of the pandemic in early 2020, with the
insufficient understanding of the virus’s origins and severity functioning as a fertile
breeding ground for misinformation to form. Some of these initial deceptive efforts
were directly launched by Russia in an attempt to replicate their success in dividing the
American public during the 2016 election, with these foreign actors seeking to “capitalize
on the fear and confusion surrounding the Covid-19 pandemic by actively promulgating
conspiracy theories.”\textsuperscript{57} In contrast to the specific targeted messages required for the
2016 influence operations, this new disinformation campaign was able to take a more
generalized strategy that merely sought to overwhelm social media sites with waves
of contradictory information and conspiratorial critiques of public health responses.
This approach fell in line with Russia’s traditional “preexisting playbook of using crises
to inflame tensions in foreign societies,” with their main focus being the politicization
of issues related to COVID-19 as a means to sow dissension rather than pushing any
particular viewpoint about the nature of the pandemic.\textsuperscript{58}

The effects of disinformation carried out by foreign actors were further compounded
by domestic social media users sharing and interacting with these conspiracy theories

\textsuperscript{56}Sam Schechner, Jeff Horwitz, and Emily Glazer. “How Facebook Hobbled Mark Zuckerberg’s Bid to Get
facebook-mark-zuckerberg-vaccinated-11631880296.

\textsuperscript{57}Julian E. Barnes. “State Dept. Traces Russian Disinformation Links”. In: The New York Times (Aug. 5,
2020). URL: https://www.nytimes.com/2020/08/05/us/politics/state-department-russian-
disinformation.html.

\textsuperscript{58}Ben Dubow, Edward Lucas, and Jake Morris. Jabbed in the Back: Mapping Russian and Chinese Information
of their own volition. These individuals effectively fill the role of what Russian strategists refer to as “useful idiots,” or general populations that “regurgitate Russian themes and ‘facts’ without necessarily taking direction from Russia or collaborating in a fully informed manner.”

By leveraging these independent third parties as a vital component of the disinformation pipeline, the originators of a piece of fake news or misleading report are able to achieve a far greater reach than would otherwise be possible if relying solely upon the strength of their internally-controlled networks. The impact of the rapid spread of misinformation can also be exacerbated by the modern inversion of media, in which reports from traditional news outlets are frequently shaped by developments on social platforms which can lend an unearned degree of legitimacy to blatant misinformation. Experimental research has supported the existence of such a link between the conventional and digital realms of media, as demonstrated in a 2021 study that showed that “people’s trusted news sources are correlated with their belief in COVID-19 misinformation” and that “larger shares of those who trust COVID-19 information from leading conservative news sources believe misinformation.”

These results call attention to the dangers of social media polluting the overall domain of public discourse, with these repercussions undermining trust in institutional sources and inhibiting the ability to maintain a shared national conception of truth.

In addition to impacting national news outlets, these “useful idiots” can also take the form of either a large body of low-level users collectively sharing similar opinions or a smaller selection of high-profile figures that espouse misinformation. The latter group have been represented during the COVID-19 pandemic by a group of 12 activists and entrepreneurs deemed the “Disinformation Dozen” in a report by the Center for Countering Digital Hate (CCDH), which determined that “65% of the shares of anti-vaccine misinformation” across Facebook, Instagram, and Twitter can be directly attributed

---

59Weisburd, Watts, and Berger, *Trolling for Trump: How Russia is Trying to Destroy Our Democracy*. 
to this group. However, in spite of the high degree of visibility of these prominent accounts, the CCDH found that these platforms “fail to act on 95 percent of the Covid and vaccine misinformation reported to them,” and further demonstrated evidence that “Instagram’s algorithm actively recommends similar misinformation.” These trends serve to highlight the substantial influence that even a small number of accounts can exert through consistent messaging and the support of a wider population of like-minded individuals, with the structure of social networks and their recommendation algorithms serving to reinforce the inherent biases of these groups within the context of a digital echo chamber.

It is critical to note that even in spite of the attempts made by mainstream social media platforms to reduce the spread of falsehoods related to COVID-19, their responses clearly failed to weaken the most pervasive trends and these misinformation campaigns still had a directly measurable impact on the spread of the pandemic. The majority of modifications made by platforms such as Facebook and Twitter were restricted to “soft remedies, such as providing a label or warning when users liked a page found to repeatedly share misinformation,” with these companies only sparingly relying upon more aggressive moderation tools to reduce the visibility of posts or remove them. The lack of a stronger response was found to have a tangible effect on public perception of the safety of the COVID-19 vaccines, with statistical analyses concluding that there exists a “substantial relationship between foreign disinformation campaigns and declining vaccination rates.” These effects were found to occur at both a national level and on a more local scale as was demonstrated by a team of researchers at the University of

---


64 Steven Lloyd Wilson and Charles Wiysonge. “Social media and vaccine hesitancy”. In: BMJ Global Health 5.10 (2020). DOI: 10.1136/bmjgh-2020-004206. eprint: https://gh.bmj.com/content/5/10/e004206.full.pdf. URL: https://gh.bmj.com/content/5/10/e004206, p. 1.
Wisconsin-Milwaukee, who relied upon mapping tools to identify a “direct correlation between locations where Twitter misinformation originated and subsequent spikes in COVID-19 infections and deaths in those areas weeks later.”\(^{65}\) The results of the research investigations discussed throughout this section accentuate the extent to which uncontrolled disinformation on social media can yield a concrete impact on human lives, emphasizing the moral consequences of failing to enact sufficiently strong protections against the malicious influence of both foreign and domestic actors.

Chapter 4

The Facebook Files: An Inside Look at Facebook’s Failures

4.1 Background on the Leaks

In spite of the frequent scandals that Facebook has faced over the course of their existence and particularly since the 2016 presidential election, information regarding their corporate misconduct has primarily come from outside sources or external analyses of the impacts of their platform on society. In line with many of the largest firms in the social media domain and the tech industry at large, Facebook has historically existed as a secretive company, closely guarding the details of the algorithms that control content on the site and providing little transparency into the executive decisions that influence the overall direction of their platform. However, this established secrecy was shattered in September 2021 with The Wall Street Journal’s release of a series of highly damaging disclosures titled “The Facebook Files”. This investigation was based upon a large collection of leaked internal documents that provided the Journal with an unprecedented degree of insight into the inner workings of Facebook, allowing them to conclude that the corporation and its subsidiary platforms “are riddled with flaws that cause harm, often in ways only the company understands” and that “despite congressional hearings, its own pledges and numerous media exposés, the company didn’t fix them.”

---

4.2 XCheck and the Inequity of Site Rules

These revelations were later revealed to have originated from Frances Haugen, a former product manager and member of Facebook’s Civic Integrity team, who chose to resign from the company and come forward as a whistleblower after witnessing the corporation’s consistent pattern of neglect and indifference towards issues of user safety and the public good.\(^{67}\) Following her first public appearance in an interview with “60 Minutes” in October 2021 and her testimony before the U.S. Senate Commerce Committee, Ms. Haugen continued to provide more documents to The Wall Street Journal and other news outlets, resulting in an extended body of reporting known as the “Facebook Papers.”\(^{68}\)

In an effort to highlight the extent of the ethical abuses carried out by Facebook, this chapter will examine the most significant leaks to illustrate the company’s failures to police disinformation and hate speech, consistently apply their code of conduct to high-profile users, and protect global users from the threats of totalitarian governments seeking to censor or persecute citizens for their actions. A secondary aim of this chapter is to call attention to the excessive power held by Mark Zuckerberg and a select few executives to unilaterally dictate the direction of the corporation’s response to crises and halt company initiatives to improve or alter the most problematic aspects of the platform.

4.2 XCheck and the Inequity of Site Rules

The very first leak in The Wall Street Journal’s original Facebook Files series directly refuted the notion that Facebook’s rules are fair and equally applied to all users of their platform, calling into question the company’s purported stance of neutrality towards content. Although Facebook has always maintained that their site policies and terms of use are consistently enforced across their entire site and that no individual is above their jurisdiction, the Journal’s disclosure of the “XCheck” or “cross check” program revealed

---

\(^{67}\)Cat Zakrzewski and Cristiano Lima. “Former Facebook employee Frances Haugen revealed as ‘whistleblower’ behind leaked documents that plunged the company into scandal”. In: The Washington Post (Oct. 4, 2021). URL: https://www.washingtonpost.com/technology/2021/10/03/facebook-whistleblower-frances-haugen-revealed/.

a set of confidential procedures for the special treatment of a set of privileged users.\textsuperscript{69} While the program was found to be completely internal to Facebook and users were not made aware of their status, the application of XCheck privileges could allow them to be “whitelisted” and blocked from any punishment for actions that violate the rules of the site, including excusing the user from the threat of having their posts taken down by moderators.

Although the final version of XCheck exposed by The Wall Street Journal clearly presented a massive threat to supposed impartiality of Facebook and provided significant potential for abuse, it is important to note that the original intention of the program was far more limited. The ability for any user of a social media site to report other users’ posts raises the problem that certain categories of users such as politicians or journalists will frequently be subject to targeted attacks and false reports from trolls or dissenting individuals, and therefore Facebook intended to provide an additional layer of caution when handling actions taken against these high profile accounts. However, while this initial aim may have been benevolent, the implementation of XCheck quickly fell prone to corruption and a lack of consistent regulations for what qualifies an account to be enrolled into the program; by 2020 “at least 5.8 million users” were included in XCheck protocols, with an internal review expressing concern that “these people can violate our standards without any consequences.”\textsuperscript{70}

While Facebook’s efforts to conceal the nature and scope of XCheck are indicative of a malicious intent to hide this program from users, the extent of the company’s deception was actually far broader. According to the reports from The Wall Street Journal, Facebook also misled their own Oversight Board, which functions as an independent authority designed to maintain accountability and transparency, with the corporation falsely informing the Board that XCheck was restricted to just “a small number of decisions”


\textsuperscript{70}Horwitz, “Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That’s Exempt.”
and had little impact on daily activity. The figures released by the Journal regarding the size of the program directly contradicted this claim and actually prompted the Oversight Board to launch an investigation into Facebook’s use of the system, betraying an obvious schism between the company executives and their oversight panel.

The result of this scrutiny was a strong rebuke from the Oversight Board, which asserted that the company leaders were “not fully forthcoming” in their explanations of XCheck and had explicitly “failed to provide relevant information” that had been requested for review. A particular case of interest to the Board was how Facebook’s lack of transparency regarding XCheck had factored into the company’s longstanding protections of posts from Donald Trump during his campaign and presidency until his suspension from the platform in January 2021. The original report from The Wall Street Journal called to attention how XCheck was in fact directly responsible for Trump’s post that “when the looting starts, the shooting starts” during the May 2020 Black Lives Matter protests remaining online in spite of its clear incitement of violence; the program ultimately escalated the decision all the way to Mark Zuckerberg, who personally opted to allow the post to stay on the site. Facebook’s Community Guidelines specifically prohibit any speech on their site that “incites or facilitates serious violence” or poses a “genuine risk of physical harm or direct threats to public safety,” so their preservation of Trump’s inflammatory rhetoric stood as a deliberate divergence from their established policies. Overall, Facebook’s actions within the context of XCheck reveal the irrefutable existence of preferential treatment for specific users with the capacity for isolated executive decisions to override any of the platform’s terms of use, undermining

---


74 Horwitz, “Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That’s Exempt.”

the corporation’s façade of neutrality.

4.3 Treatment of Political Content After 2016

While the extent of Facebook’s influence over the 2016 presidential election has been extensively documented and is described in detail in Chapter 3, their role in handling public discourse during the subsequent 2020 presidential election has been far less publicized. While the company provided vague public relations statements expressing a commitment to combating fake news and reducing the prevalence of disinformation on their platform in the wake of the 2016 election, the exact nature of how they aimed to accomplish this remained unclear. This equivocal stance was maintained by Mark Zuckerberg in his April 2018 hearing before the Senate Commerce and Judiciary Committees, in which he failed to provide any specific outline of a plan of response but conceded that “it’s clear now that we didn’t do enough to prevent these tools from being used for harm, as well. And that goes for fake news, for foreign interference in elections, and hate speech… we didn’t take a broad enough view of our responsibility, and that was a big mistake.”

Facebook sought to provide themselves with a diverse range of possible responses to these admitted deficiencies through the establishment of a range of “Integrity Teams” designed to research specific social issues on the platform and develop potential solutions, yet these teams frequently found the results of their work rejected by executive decision. One particularly influential figure in this pushback has been Vice President of Global Public Policy Joel Kaplan, who has frequently stepped in to provide a unilateral rejection of changes proposed by these teams that could have the effect of limiting conservative speech or reducing the reach of right-wing news outlets on the site. While Kaplan is not the only executive within Facebook with the power to directly control the implementation

---


of specific features and dictate the overall direction of the platform, the ability for a single individual to have this degree of unchecked influence presents a major threat to the legitimacy of the company’s purported commitment to reform.

The full picture of Facebook’s response to address their platform’s critical flaws only became clear following the release of the Facebook Files, which revealed that the company approved the development and deployment of two algorithmic tools aimed at limiting the presence of misleading content on their site. The first, known as “Sparing Sharing,” served to prevent the promotion of content from “hyperposter” accounts that aim to share illegitimate information to a large user base at alarming rates. This feature had actually been rejected by Kaplan’s review when it was initially proposed, but was later implemented in a diminished form. While this measure was targeted at problematic accounts, the second modification, dubbed “Informed Engagement,” sought to minimize the spread of posts and linked sources that users shared without reading themselves and to directly reduce the proliferation of low quality news links with sensationalist headlines. In combination, this pair of revisions to the site’s algorithms were able to effectively shrink the presence of harmful disinformation on the platform. However, after an internal “political ideology analysis” study deemed that these changes had the result of “suppressing the traffic of major far-right publishers,” the company chose to repeal the Informed Engagement program and return to their prior stance of leniency towards all news sources.

While these tools served as the culmination of a years-long effort towards the general improvement of the platform, Facebook also moved to prepare and enact stricter measures for the period immediately preceding the 2020 election in an effort to limit their direct impact on the electoral results and protect themselves against being held accountable for any new attempts at foreign interference. As part of the rollout of this

---


79Hagey and Horwitz, “Facebook’s Internal Chat Boards Show Politics Often at Center of Decision Making”.

80Hagey and Horwitz, “Facebook’s Internal Chat Boards Show Politics Often at Center of Decision Making”.

new plan of electoral strategies, Zuckerberg professed the company’s “responsibility to protect our democracy” and announced a sitewide halt on political advertising in the week leading up to the election, the creation of a new scheme of informational labels to flag potentially misleading posts, and a temporary reduction on the rate at which users could create new Facebook groups and engage in the mass sharing of posts.\textsuperscript{81}

Although many of these modifications seem minimal they still presented a notable shift in policy for a corporation that has historically resisted changes that could detract from the levels of user engagement on their platform. An unnamed former Facebook executive admitted that the objective of this novel approach “was to make sure Facebook wasn’t the story” of the election, attempting to avoid a scandal similar to 2016.\textsuperscript{82} The company also placed an internal emphasis in the leadup to the election on establishing “break glass” strategies intended to provide a framework for controlled intervention in the case of a contested result, yet instead of putting these schemes into action, leaks revealed that Facebook “began relaxing its emergency steps in November.”\textsuperscript{83} In spite of their extensive planning efforts, they ultimately opted to withdraw these specialized measures as soon as the election was decided, underlining the company’s unwillingness to commit fully to functional solutions for the problem of disinformation.

4.4 Influence of Facebook on the January 6th Insurrection

While Facebook reportedly considered their work during the 2020 election to be a clear success, their premature declaration of victory and the abrupt truncation of their heightenened security measures opened the door for a rampant wave of disinformation questioning the results of the election. Facebook’s immediate return to normalcy despite these looming dangers reflected their general reluctance to deviate from the company’s status

\textsuperscript{81}Mark Zuckerberg. The US elections are just two months away, and with Covid-19 affecting communities across the country, I’m concerned about.. Sept. 3, 2020. URL: \url{https://www.facebook.com/zuck/posts/10112270823363411}.

\textsuperscript{82}Craig Timberg, Elizabeth Dwoskin, and Reed Albergotti. “Inside Facebook, Jan. 6 violence fueled anger, regret over missed warning signs”. In: The Washington Post (Oct. 22, 2021). URL: \url{https://www.washingtonpost.com/technology/2021/10/22/jan-6-capitol-riot-facebook/}.

quo of perceived neutrality and avoid accusations of political bias, particularly from conservative politicians and public figures who were quick to accuse the platform of silencing them for their beliefs.

On the morning of January 6th, Facebook detected an unprecedented spike in hate speech and incitements of violence, with “user reports of ‘false news’ hitting nearly 40,000 per hour” and “the account reported most often for inciting violence was @realDonaldTrump—the president’s official account.” In an effort to staunch this unrelenting wave of rule-violating content, employees were instructed to reintroduce many of the “break glass” measures that had been used during the election, yet while some were successfully put into effect, “other measures, such as preventing groups from changing their names to terms such as Stop the Steal, were not fully implemented because of last-minute technology glitches.” The inadequacy of these attempts to mitigate the spread of this incendiary content on the day of the riots and Facebook’s struggles to deploy fully operational tools for intervening during crises stands as a major failure for both the company and their corporate strategy.

In addition to their repeal of some of the technological features that had boosted the platform’s defenses against disinformation during the end of the campaign and the election, Facebook had also experienced a mass exodus of staff who “took leaves of absence or moved on to other jobs” after it became clear that their electoral work would be reversed. The company also moved to dismantle their Civic Integrity team after executives “had grown weary of the team’s criticisms of the company.” This vital group, whose members included the future whistleblower Frances Haugen, had been frequently referenced throughout internal reports that suggested that “leadership was not happy with how loud the Civic Integrity team was…they wanted to decentralize the team and

---

84 Timberg, Dwoskin, and Albergotti, “Inside Facebook, Jan. 6 violence fueled anger, regret over missed warning signs”.
85 Timberg, Dwoskin, and Albergotti, “Internal Alarm, Public Shrugs: Facebook’s Employees Dissect Its Election Role”.
86 Timberg, Dwoskin, and Albergotti, “Inside Facebook, Jan. 6 violence fueled anger, regret over missed warning signs”.
87 Timberg, Dwoskin, and Albergotti, “Inside Facebook, Jan. 6 violence fueled anger, regret over missed warning signs”.

the power they had."\textsuperscript{88} By choosing to eliminate the Civic Integrity team in spite of the ongoing disinformation crisis, Facebook fundamentally staked their position of only protecting public interests when absolutely necessary to the reputation of the company, and in all other circumstances would return to a baseline of minimal protections and oversight.

Due to this internal strife and the lack of a consistent plan for approaching the novel threats posed by an angry and energized population of Trump voters who believed their candidate had been cheated out of the election, Facebook found themselves incredibly vulnerable to the pressure of this incited base. The limitations of the platform’s traditional systems for filtering out disinformation and misleading political content were clearly on display through the entire period between the election and January 6th, with Facebook admitting in an internal report that “Stop the Steal (StS) grew rapidly after the election as a movement, but our enforcement was piecemeal.”\textsuperscript{89} Although the company employed their “exceptional measures” to institute a permanent ban on the original Stop the Steal group within 24 hours of its creation on election night, it failed to extend this decision to many of the numerous groups that sprung up in its place over the following two months and ultimately left them online until after the attack against the Capitol took place on January 6th.\textsuperscript{90}

However, the scale of the rapid expansion in post-electoral disinformation was far more extensive than Facebook’s admission implied; a leak to The New York Times indicated that by November 9th, 2020, just two days after Joe Biden was declared the winner of the election, “as much as one out of every 50 views on Facebook in the United States, or 10 percent of all views of political material, was of content declaring the vote fraudulent.”\textsuperscript{91} These staggering figures highlight the ubiquitous reach of posts aimed

\textsuperscript{88}Timberg, Dwoskin, and Albergotti, “Inside Facebook, Jan. 6 violence fueled anger, regret over missed warning signs”.


\textsuperscript{90}Timberg, Dwoskin, and Albergotti, “Inside Facebook, Jan. 6 violence fueled anger, regret over missed warning signs”.

\textsuperscript{91}Mac and Frenkel, “Internal Alarm, Public Shrugs: Facebook’s Employees Dissect Its Election Role”.
at eroding confidence in the American electoral process and its results, and completely undermine any claim on the part of Facebook that they had done their due diligence towards upholding a heightened state of security until the inauguration of the next president. Rather than maintaining a commitment throughout the entire transition of power, Facebook instead opted to overlook these blatant signs of unrest and enabled many of the groups that organized and participated in the insurrection to coordinate on their platform.

4.5 Violence and Censorship in Developing Countries

While Facebook’s corporate decisions and the application of their site rules within America are evidently inconsistent, these issues are not only domestic and are particularly exacerbated and can have even graver consequences when shifting focus to their growing markets in developing nations. In fact, a combination of the company’s limited resources for moderation in other languages and a purposeful strategy of prioritizing rapid market expansion over user safety and fair enforcement has led to a surge of unregulated dangerous activity on their platform. The contrast between Facebook’s strategies in developing and established markets was directly explored within the Facebook Files leaks, implicating the company leadership in these deliberate actions.

In many cases this corporate failure is driven chiefly by a shortage in global moderation staff, with documents provided to The Wall Street Journal as part of the Facebook Files demonstrating that “in some countries where Facebook operates, it has few or no people who speak the dialects needed to identify dangerous or criminal uses of the platform”.\(^{92}\) While it is obviously difficult to be able to find and hire qualified moderators to cover every possible language that could be spoken on the platform, Facebook’s clear inaction and inability to mitigate this problem stands as a major act of negligence towards the users that speak these underrepresented languages; as a result, users in these

markets will ultimately be engaging with each other on an effectively unmoderated site that lacks the means to enforce their established community standards in any consistent manner. Without the means to interpret the content that users provide in these languages, Facebook ultimately has no way of discerning whether a specific post violates any rules (except for photo or video content that unmistakably infringes upon some visual guideline), and therefore leaves some developing countries prone to all manner of abuses being carried out on the platform. This shortage of staff and translators for many of the dialects spoken on Facebook also presents an algorithmic challenge as the company will have no capability to develop or test their technical tools to account for posts in these languages, such that the “artificial-intelligence systems that form the backbone of Facebook’s enforcement don’t cover most of the languages used on the site.”

However, as disclosed in the Facebook Files, this phenomenon of lacking moderation in developing nations also extends to areas of the world where Facebook does possess sufficient moderation capabilities, reflecting an intentional plan to turn a blind eye to malicious conduct in these burgeoning markets to avoid quashing their rapid expansion. Although the expansion of Facebook within Western nations has largely stagnated as the platform has become “saturated” in these established markets, executives have placed a particular focus on their business in many areas of Asia and Africa, where public internet connectivity continues to grow more widespread and opens up the potential for massive waves of new users to join the platform.

Even in cases where Facebook leaders are fully aware of the nature of the online discourse in its developing markets and the tangible ramifications this content could have for its users’ digital and physical safety, the company has opted to treat this as “simply the cost of doing business” according to a former VP in global marketing.

The sources provided to The Wall Street Journal spotlighted numerous examples of

---

93 Scheck, Purnell, and Horwitz, “Facebook Employees Flag Drug Cartels and Human Traffickers. The Company’s Response Is Weak, Documents Show.”


95 Scheck, Purnell, and Horwitz, “Facebook Employees Flag Drug Cartels and Human Traffickers. The Company’s Response Is Weak, Documents Show.”
the types of vile conduct that employees regularly discover on the platform, including reports that “a Mexican drug cartel was using Facebook to recruit, train, and pay hit men,” “human traffickers in the Middle East used the site to lure women into abusive employment situations,” and “armed groups in Ethiopia used the site to incite violence against minorities.” However, in spite of the employees possessing concrete proof of all of these atrocities and demonstrating clear grounds for their removal from the platform, the majority of these complaints were ignored entirely or resulted in only a few pages being taken down while leaving the bulk of the problem unaddressed. Despite the vast majority of Facebook’s user base residing in foreign markets the company commits most of its enforcement resources to domestic issues, with a leaked statistic that “only 13%” of moderation time is spent on cases outside of the United States substantiating how the company consistently prioritizes “brand safety” over a genuine concern for social issues and the wellbeing of their users.

This inexcusable disregard was particularly evident within Facebook’s handling of the instigation of violence and hate speech against the Rohingya ethnic group in Myanmar, with the company admitting that it had served as an “‘enabling environment’ for the proliferation of human rights abuse”. At the center of this devastating crisis was a “systematic campaign” run by the Myanmar military seeking to “generate widespread feelings of vulnerability and fear that could be salved only by military’s protection,” and this disinformation scheme helped the totalitarian regime strengthen their control over the nation and defend their actions against the Rohingya. The corporation’s degree of complicity in this ethnic cleansing was expressed even more strongly by a second anonymous whistleblower who submitted additional documents to the SEC in the wake of Ms. Haugen’s public statements, with the unnamed leaker stating that “I, working for
Facebook, had been a party to genocide.”

While most of these examples of Facebook’s gross neglect are linked to their traditional stances of nonintervention and lax moderation, the company has demonstrated a willingness to institute specialized censorship measures on their platform in order to remain online in certain expanding markets. A leak to The Washington Post detailed that in 2020 the Vietnam Communist Party directly requested that Facebook “censor anti-government dissidents or risk getting knocked offline,” and in spite of his traditional stance of upholding free discourse “Zuckerberg personally decided that Facebook would comply with Hanoi’s demands.” As a result of this executive decision, Facebook “significantly increased censorship of ‘anti-state’ posts, giving the government near-total control over the platform” and surrendering an unprecedented level of power to outside actors in contrast to the company’s usual stringent influence over every facet of their platform. When all of these global issues are taken in conjunction, they present an incriminating picture of Facebook as being inconsistent and ill-prepared for the pressures of managing issues of human rights and international politics, with the corporation rejecting any fixed moral or ethical position in favor of unchecked expansion.

4.6 Racial Bias and Hate Speech

While issues of disinformation and deceptive political content continue to be one of the most critical challenges for Facebook to solve within both domestic and foreign markets, a related issue that can have a pervasive impact is the widespread presence of hate speech and discrimination. The public awareness of Facebook’s inability to adequately police hate speech began long before the Facebook Files leaks, with the company launching a two-year “Civil Rights Audit” in 2018 under pressure from prominent civil rights leaders

---


102 Dwoskin, Newmyer, and Mahtani, “The case against Mark Zuckerberg: Insiders say Facebook’s CEO chose growth over safety”.

---
4.6. Racial Bias and Hate Speech

and members of Congress to assess the platform’s equitability. While Facebook intended for the audit to stand as evidence of their stated commitment to fighting discrimination, the company found themselves facing heavy criticism from the report, which concluded that they had consistently made “vexing and heartbreaking decisions” in this domain and that its continued protection of posts from Trump during his presidency resulted in “significant setbacks for civil rights.”

Furthermore, the auditors concluded that “Facebook’s approach to civil rights remains too reactive and piecemeal” and highlighted how “many in the civil rights community have become disheartened, frustrated and angry after years of engagement where they implored the company to do more to advance equality and fight discrimination.”

In addition to the Civil Rights Audit’s censure of Facebook’s existing policies and treatment of discriminatory content, the report also outlined a wide range of suggestions for possible changes or additions that could be made to the platform’s rules and technologies to improve their response to civil rights issues. Even in the face of this direct censure the company still pushed back against many of the stronger or more intensive recommendations submitted by the auditors, as Chief Operating Officer Sheryl Sandberg argued in an equivocal statement that “while we won’t make every change they call for, we will put more of their proposals into practice.”

Ultimately, a year of intense outside pressure and the establishment of a Civil Rights Team to oversee the implementation of these proposals resulted in some considerable progress, with the VP of the new team Roy L. Austin announcing in November 2021 that out of the “117 actions and recommendations” laid out by the audit, “65 have been implemented, and 42 are either in progress or are ongoing.” Unlike the dissolution of the Civic Integrity team following the 2020 election, the enduring existence of this Civil Rights team has had a

---


notable impact in forcing Facebook to consider social issues in their decision-making and has functioned as a check against the company’s frequent negligence.

However, despite this team’s concerted efforts, executive decisions have continued to obstruct some of the most significant improvements that could be made to reducing the prevalence of hate speech on the platform. One such decision was disclosed in a leak to The Washington Post, which revealed how Facebook had abandoned their extensive “Worst of the Worst” program for detecting and removing racial bias and hate speech. Historically, Facebook’s detection and treatment of hate speech has followed what the company refers to as a “race-blind” approach, under which any negative statement directed at a specific population is regarded as equally severe, regardless of whether it was addressed towards a group that was “more likely to be targets of hate speech versus those that have not been historically marginalized.” The direct consequence of this policy caused “comments like ‘White people are stupid’ [to be] treated the same as anti-Semitic or racial slurs.” Facebook’s race-blind stance also had the side effect of censoring a significant number of legitimate posts discussing systemic forms of discrimination, because “challenging white supremacy and White men is an important part of having dialogue about racism.”

In an effort to address these major flaws within their strategy for handling hate speech, Facebook researchers moved to institute a complete revision of their automated systems for filtering discriminatory content, with the WoW project placing a heightened emphasis on punishing hateful speech towards the most heavily targeted minority populations, including users who are “Black, Jewish, LGBTQ, Muslim or of multiple races.” This shift in focus would have provided the platform with a more nuanced and responsible

---

108 Dwoskin, Tiku, and Kelly, “Facebook to start policing anti-Black hate speech more aggressively than anti-White comments, documents show”.
109 Dwoskin, Tiku, and Kelly, “Facebook to start policing anti-Black hate speech more aggressively than anti-White comments, documents show”.
scheme for protecting their most vulnerable users from discrimination, falling in line with their professed commitment to improve their treatment of civil rights following the results of the 2020 audit. However, company executives led by Joel Kaplan strongly rebuffed the proposal on the basis of a fear of negative reactions from “conservative partners,” who they anticipated would argue that the restructured algorithms violated their freedom of speech; a leaked internal document from Kaplan’s team explicitly stated that some of these partners contend that “hate targeted toward trans people is an expression of opinion,” underscoring the company’s full awareness and complicity in perpetuating the discriminatory conduct on their platform.\footnote{Dwoskin, Tiku, and Timberg, “Facebook’s race-blind practices around hate speech came at the expense of Black users, new documents show”}
Chapter 5

Moderation and Recommendation Algorithms

5.1 The Uses of Algorithms on Social Media Platforms

While many of the social abuses carried out on social media platforms can be traced back to distinct executive decisions or blatant negligence by their respective companies as described in Chapters 3 and 4, a certain degree of focus must also be placed upon the role of algorithms in contributing to the harmful climate of social media. The use of algorithmic technologies within social media sites is ubiquitous, with almost every facet of these platforms depending upon some form of automated system designed to display content to the user, provide recommendations for pages that they may enjoy, filter out harmful posts from being viewed by the public, and deliver the targeted advertisements that serve as a primary source of revenue. While all of these functions are important to the overall operation of a social media site, the moderation and recommendation systems that control the existence and spread of content on their platforms are by far the most critical in terms of influencing the direction of site discourse. When these algorithms are implemented without considering all of the possible means by which they could be circumvented or exploited, the platform ultimately makes itself vulnerable to the threats of disinformation, hate speech, and other forms of pervasive content manipulation, endangering both their user population and society at large as a consequence.

Although major social media platforms typically treat the implementation details of
their algorithms as critical company secrets, a combination of transparency measures prompted by external pressure and leaked information from internal documents such as the Facebook Files have provided a significant degree of insight into the inner workings of these sites and the attitudes of these corporations towards algorithmic intervention.\textsuperscript{112} In addition, while these companies often provide reports and statistics promoting their own algorithms and making strong claims about their effectiveness, leaked data has frequently indicated that these statistical measures are often based on highly misleading or manipulated results that conceal the platforms’ technical limitations. In some of these cases, it is important to consider whether these misrepresentations constitute examples of deliberate false advertising or misleading reports to stockholders and therefore could be used to prosecute the offending companies for these intentional violations. Moreover, efforts towards transparency will directly aid users in gaining a clearer understanding of the technical shortcomings of these sites and help them to better protect themselves in situations where automated moderation features are lacking.

This chapter contains an analysis of how these respective platforms portray their use of moderation and recommendation algorithms and the reality of how they work in practice. By exploring this dichotomy between social media companies’ claims and their actual implementations, it is possible to identify areas where these corporations could improve their algorithmic approaches. This discussion places a particular emphasis upon identifying the specific technical components of these algorithms that are responsible for the inadequacy of their current implementations, effectively creating a roadmap of features that could be improved or restructured. After this broad technical overview, this chapter concludes by examining an additional series of leaks from the Facebook Files not previously mentioned in Chapter 4. These highlight Mark Zuckerberg’s prominent role in directly controlling major design choices for the platform including the creation of a novel statistical metric that forms the backbone of the current version of the site’s recommendation algorithms.\textsuperscript{113} This specific example calls attention to the troubling

\textsuperscript{112}The Wall Street Journal, \textit{The Facebook Files: A Wall Street Journal Investigation}.

trend within some of these companies of allowing for executives to take the power out of
the hands of developers and make unilateral decisions about technical features of their
platforms, effectively negating the expertise of their development teams and allowing
for risks to be taken with minimal oversight.

5.2 The Legal Standing of Moderation

While moderation algorithms may appear to be one of the most essential components
of constructing a safe internet platform, it is important to recognize that they have not
actually been a part of social media since its inception and have instead grown into use
during the early 2010s in an effort to protect their platforms from liability and safety
issues. Although these moderation tools have been developed in conjunction with
their platforms’ existing human moderation services and were originally intended to
supplement their traditional content review strategies rather than replace them, the
importance of automated moderation tech has continually grown. The complexity of
these systems and their platforms’ degree of reliance upon them has increased rapidly
since their inception in order to account for the massive scale of modern platforms and
the multitude of types of content that require moderation, yet many of these companies
still maintain a conservative approach towards moderation and seek to minimize the
extent of content being removed from their services.

A critical background for any discussion of moderation on social media requires an
examination of the legal standing of user-generated content and the obligation for com-
panies to remove offensive or harmful material from their sites. Although social media
obviously has a global audience, the fact that many of the world’s largest social platforms
are based in America entails that they are primarily bound to U.S. public speech and
internet laws. As a result, companies such as Facebook, Instagram, and Twitter have
effectively structured their entire approaches to content moderation around American

ideals of free speech and an open internet. However, their expansion into the international sphere has led to conflict in territories that promote censorship or place strong limitations upon public discourse. It is critical to note that while the First Amendment to the U.S. Constitution asserts that “Congress shall make no law... abridging the freedom of speech” it does not place any restrictions upon private companies exercising the power to dictate the forms of speech that are acceptable within their jurisdiction. Therefore, American social media platforms are currently allowed to define their own rules for online communication and are effectively free to engage in any degree of moderation or censorship of content on their sites, with the U.S. Supreme Court repeatedly upholding the position that “merely hosting speech by others is not a traditional, exclusive public function and does not alone transform private entities into state actors subject to First Amendment constraints.”

Beyond the ramifications of the First Amendment, social media platforms are provided with an additional layer of legal protection under Section 230 of the 1996 Communications Decency Act, which laid the foundation for the American legal system’s treatment of content in cyberspace. The key clause of Section 230 stated that “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,” which had the effect of directly exempting internet services from liability for any content uploaded or posted by their users. This provision was ultimately fundamental to the development of the present-day cyberspace, as without its overarching protection it would have been completely impractical for any internet service to provide the capacity for large-scale user-generated content, which would have required manual review to prevent the host from being found culpable for any illegal or indecent content. As a result, the creation of any extensive social media network would have been legally infeasible, along with the

---

development of any services such as Google Drive, iCloud, or Dropbox that are used to host and share user-provided files.\textsuperscript{118}

While Section 230 has been critical to the evolution of the modern internet, it has also allowed for social media companies to adopt a distinctly lenient approach towards handling content moderation on their platforms. In general, the major corporations within this burgeoning domain in the late 2000s opted to forgo extensive moderation efforts under the guise of maintaining stances of neutrality and championing free speech for the world. The underlying motive for this approach was to maximize the potential for the platform’s growth by allowing for all content to flourish unless it was explicitly illegal, yet the companies would occasionally demonstrate the capacity for more robust moderation when it proved beneficial to their global expansion. This was epitomized by contemporary accounts of the stricter content guidelines that Google instituted in 2008 on YouTube in foreign nations in order to comply with censorship requests.\textsuperscript{119} This lax moderation style was even more relevant in the case of Facebook in the company’s early years, as their platform became globally available to all users outside college campuses in 2006 yet did not publish any kind of public community guidelines until 2011 and instead treated all moderation decisions on a case-by-case review basis.\textsuperscript{120}

\subsection*{5.3 Algorithmic Moderation Techniques}

Before social media firms began moving to implement automated moderation programs in the 2010s, all of their early efforts in this domain were composed entirely of human moderation employees.\textsuperscript{121} These teams were originally tasked with manually inspecting posts to determine whether they contained illegal content or judge whether they had a damaging effect on the users of the platform or society at large, but today their role is

\begin{itemize}
  \item \textsuperscript{120}Facebook. \textit{Facebook Community Standards}. Jan. 27, 2011. URL: \url{https://web.archive.org/web/2010127224041/https://www.facebook.com/communitystandards/}.
  \item \textsuperscript{121}Checkstep. \textit{The Evolution of Content Moderation Rules Throughout The Years}. Medium. Apr. 1, 2021. URL: \url{https://medium.com/checkstep/the-evolution-of-content-moderation-rules-throughout-the-years-bccc9859cb31}.
\end{itemize}
supplemented with the use of algorithmic tools. Scholars in the field have highlighted how these “decisions were made reactively in response to signals that were reported to companies through media pressure, civil society groups, government, or individual users” rather than identifying an exact set of moderation steps to provide a comprehensive solution to the challenges faced by their platform.\textsuperscript{122} Without a strong internal drive to institute more extensive systems for managing content, these companies will consistently fall prone to insufficient moderation and will place an unmanageable degree of responsibility upon their remaining employees to engage in removing posts by hand.

In an effort to address the many moderation challenges described in Section 5.2 and bolster the limited resources of their human moderation teams, many social media companies began to introduce artificial intelligence (AI) and machine learning (ML) tools to improve their platforms. Unfortunately, many of these AI and ML systems are insufficiently understood because they function as “black boxes” that inhibit the developers from adequately predicting and controlling the outputs of these algorithms.\textsuperscript{123} Therefore, these technologies can frequently have unanticipated harmful consequences for social media platforms and their massive user populations, with inaccurate moderation technologies having a particularly deleterious influence due to the potential to censor legitimate speech or allow users to be exposed to vile and discriminatory content. The only way to resolve this ongoing dilemma is through the implementation of “AI explainability” concepts, which attempt to develop mechanisms to clarify the AI decision-making process and provide a legible record to assess how the system erred when inaccurate judgments are made.\textsuperscript{124}

While many of the more complicated automated AI moderation tools struggle to achieve consistent results, it is important to note that certain simpler technologies such as file hash matching systems have achieved a high degree of sophistication and have a wide range of practical uses. Hash matching is typically used for identifying illegal content or


detecting cases of copyright infringement by computing “a unique digital ‘fingerprint’” for each file in a database that new uploads will be checked against, such that a violating post can be identified by the presence of a matching hash.\textsuperscript{125} While a traditional hashing approach operates on the binary data of the file and can be undermined through any small alteration to its content, modern hash matching technologies such as Microsoft’s PhotoDNA service provide robust hash implementations that preserve the root content of the file and are heavily resistant to visual edits and manipulations.\textsuperscript{126} Similar techniques are employed within Google’s proprietary ContentID technology used to detect instances of video and audio copyright violations on YouTube, which allows “YouTube users to create digital hashes for their video content” that will then be checked against every video uploaded to the platform to immediately identify the presence of even fleeting snippets of copyrighted material with high accuracy.\textsuperscript{127}

Despite their limitations of relying too heavily upon automated tools for moderation, modern social media sites depend upon these technologies almost exclusively for evaluating the compliance of new posts with their community standards or terms of service. There are two primary categories of artificial intelligence that are ubiquitous throughout current platforms, with image recognition algorithms being used to analyze visual media and natural language processing (NLP) frameworks allowing the moderation system to interpret the meaning of a user’s textual post.\textsuperscript{128} Although image recognition systems offer the potential to incrementally improve detection capabilities by incorporating data from the decisions of human moderators to refine their algorithmic predictions, these approaches are incapable of integrating “nuanced and contextual insights into their detection procedures” and as a result have historically been limited in their ability to provide effective moderation.\textsuperscript{129} Furthermore, the effectiveness of any image recognition implementation is heavily dependent upon the datasets that it uses to train its machine

\textsuperscript{125}Cambridge Consultants, Use of AI in Content Moderation, p. 48.
\textsuperscript{127}Singh, Everything in Moderation, p. 12.
\textsuperscript{128}Singh, Everything in Moderation, p. 14.
\textsuperscript{129}Singh, Everything in Moderation, p. 14.
learning models and inform the program’s classification processes. Consequently, low-quality or biased data inputs can completely undermine the validity of the system’s outputs and its ability to correctly detect violating content.

While NLP algorithms are directed at the simpler level of deciphering the underlying meanings of text rather than complex visual content, they suffer from many of the same persistent defects of being unable to model elements of satire or the criticism of social oppression, and are further hampered by the inability to “capture the full breadth, context, and nuances of [extremist content and disinformation].” For example, as discussed in Section 4.6 many of Facebook’s existing systems for detecting hate speech experience unacceptably high error rates for both positive and negative classification, with the algorithms failing to identify some blatant forms of discrimination while over-indexing on certain kinds of benign speech. Major advancements have been made in this field in recent years through the development of theoretical tools such as sentiment analysis and word embeddings that provide the technical ability to discern the meaning and tone of sentences. Sentiment analysis consists of a collection of algorithmic strategies that evaluate the tone of a message, spanning a range of possible implementations “from simply labelling them as positive or negative, to more subtle labelling including the level of emotion.” These sentiment analysis systems are augmented by the use of word embedding preprocessing systems, which attempt to resolve the challenge of representing the immense complexity of human language by “grouping semantically similar words close to each other” in order to yield a significantly smaller feature space. However, despite the promising results achieved by these algorithms they are still limited in their practical effectiveness and struggle with interpreting more complex semantic structures such as irony or legitimate criticism of discriminatory speech.

\[^{130}\text{Singh, } Everything in Moderation, p. 15.\]
\[^{131}\text{Dwoskin, Tiku, and Timberg, “Facebook’s race-blind practices around hate speech came at the expense of Black users, new documents show”.}\]
\[^{132}\text{Cambridge Consultants, } Use of AI in Content Moderation, p. 50.\]
\[^{133}\text{Cambridge Consultants, } Use of AI in Content Moderation, p. 50.\]
\[^{134}\text{Cambridge Consultants, } Use of AI in Content Moderation, p. 6.\]
Chapter 5. Moderation and Recommendation Algorithms

5.4 A False Sense of Security from Ineffective Moderation

Although moderation is obviously one of the most vital components for maintaining a safe and stable online environment, it is imperative for any social media corporation to provide a realistic outlook on the state of their moderation systems and their ability to detect and handle harmful conduct on their platform. The risk of overreliance upon algorithmic techniques that are unable to capture many violations of the company’s community guidelines can cause them to be overconfident in their deficient moderation capabilities and not devote sufficient attention to adequately protecting the users of their platform, allowing for significant abuses to take place. As seen throughout the discussions of the Facebook Files leaks in Chapter 4 and particularly in Section 4.6, the combination of a dearth of human resources and limited algorithms can allow for all manner of illegal and offensive content to be spread on a social platform, placing a moral obligation upon social media companies to provide a more consistent commitment to their moderation.

These corporations’ shift towards a hybrid form of content review effectively established a new “status quo of content moderation” as an “endless, reactive game of ‘whack-a-mole,’” struggling to adapt to changing trends while users discovered techniques to bypass punishment. Under this new “top-down” moderation structure, social media platforms began removing and issuing warnings towards a far wider range of content than had previously been possible. However, these moderation decisions can actually be regarded as moving in the direction of decreasing transparency due to the heavy usage of black box AI tools for indiscriminately removing content and the relative lack of human staff to confirm or clarify the outcomes of these processes. In fact, Facebook did not even provide the ability for users to submit an appeal for the manual review of posts that were automatically removed by their algorithms until 2018, calling attention to the limitations of their original company policies and lack of preparation for

---


136 Griffin and Wilmore, “The Tragedy of Social Media Moderation”. 
5.4. A False Sense of Security from Ineffective Moderation

the dilemmas associated with AI moderation.\textsuperscript{137}

In addition to the problems with transparency and uncertain community standards, the threat of increasing moderation on mainstream platforms such as Facebook and Twitter has also had the unforeseen consequence of prompting banned users and followers of extremist ideologies and conspiracy theories to move to less regulated services in order to maintain their hateful and inflammatory discourse. While this exodus of users to “alt-tech” platforms has been construed as a victory for the moderation standards of the established sites, in practice it can be regarded as having a net negative influence on the overall safety of the internet by creating “deeper echo chambers for their hate, which can make identifying and tracking threats of real violence more difficult.”\textsuperscript{138} This impact has been directly illustrated by fringe political platforms such as Parler that were instrumental in the coordination of the January 6th insurrection; the lack of moderation standards and the protection of anonymity on these alternative sites poses a legitimate danger to society by giving their users license to engage in blatant violations of decency and the incitement of violence.\textsuperscript{139}

A further factor that drastically increased the usage of automated moderation systems was the COVID-19 pandemic, which prompted several of the major internet companies to test how their platforms would perform if they relied solely upon AI tools to moderate content. This shift was driven by the practical restriction that “content moderators could not work from home because of concerns about how people’s data may be handled,” but the resulting changes in content review standards reflected the critical value that human employees still hold in the moderation process.\textsuperscript{140} Facebook and Google both emphasized the increase in the total volume of removed content during the second quarter of 2020 as proof of the effectiveness of their algorithms, yet this trend likely encompassed a


\textsuperscript{138}Griffin and Wilmore, “The Tragedy of Social Media Moderation”.


\textsuperscript{140}Mark Scott and Laura Kayali. “What happened when humans stopped managing social media content”. In: \textit{Politico} (Oct. 21, 2020). URL: \url{https://www.politico.eu/article/facebook-content-moderation-automation/}. 
far larger share of algorithmic false positives rather than legitimate removals. In particular, Facebook’s Q2 2020 Community Standards Enforcement Report admitted that the prevalence of several forms of harmful content increased during the early months of the pandemic and announced the introduction of human employees back into several stages of the moderation pipeline, stating that “there will continue to be areas where we rely on people to both review content and train our technology.” Although this concession may point towards a more robust and balanced moderation strategy moving forward, the choices of these companies to engage in this “impromptu experiment” in a domain as critical as moderation encapsulates their clear lack of regard for user safety and the consequences of their corporate decisions on society.

5.5 Recommendation Technologies and Algorithmic Amplification

In contrast to the detailed legal framework surrounding the study of moderation algorithms covered in Section 5.2, the area of content recommendation lacks any clear sources of outside regulation or direction. However, the extent of recommendation systems’ effects and influence over user behavior is arguably as damaging as that of poorly implemented moderation technologies. Recommendation algorithms serve as the primary filtering mechanism for the content that users see and interact with on social media, determining which top posts will be displayed each time a user accesses the platform and defining the order in which further posts and advertisements will be displayed as they scroll through their main feed. This structure is integral to the core mechanisms through which users interact with their social media platforms, including Twitter’s Home

---

141 Scott and Kayali, “What happened when humans stopped managing social media content”.
143 Scott and Kayali, “What happened when humans stopped managing social media content”.
5.5. Recommendation Technologies and Algorithmic Amplification

timeline and Facebook’s News Feed (recently rebranded “Feed” in February 2022).\textsuperscript{145} The critical shared feature across these systems is that the posts displayed to the user are reviewed by a collection of algorithms to display a combination of content from pages that the user follows, topics that they have displayed interest or a history of interaction with, and general viral content that reaches a large proportion of the site’s audience.

Due to the fact that recommendation algorithms are responsible for curating the main content streams on their respective platforms, they have an extensive degree of influence over an individual’s usage of social media. While an effective recommendation scheme can provide users with personalized feeds that consistently engage them and prompt them to spend more time on the platform, poorly implemented systems will not display enough relevant or interesting content and may cause users to reduce their usage or even abandon the service. This engagement problem is particularly important for social media companies because of their reliance upon advertisements inserted into the user’s feed as the chief source of earnings, as exemplified by reports in 2020 from Facebook that “more than 98% of its total revenue” came from advertisers.\textsuperscript{146} Therefore, the profit of these platforms is almost directly proportional to the amount of time they can convince users to stay on the platform, resulting in the primary motives of recommendation algorithms being user retention and the delivery of maximally engaging content.

The present state of recommendation technologies on social media has arisen over the course of the past decade through a cyclical process of corporations making tweaks to their algorithms, testing the consequences through live results, and further adjusting their implementations in response to maximize profit and limit the promotion of low-quality content. Many prominent systems such as Facebook’s News Feed did not always rely upon sophisticated algorithms, with the original version of the News Feed deployed in 2006 displaying only a “personalized list of activity updates from friends” and the first ranking system not being launched until 2009 to prioritize more significant friend

\textsuperscript{145}Mitchell Clark. “Facebook rebrands News Feed after more than 15 years”. In: The Verge (Feb. 15, 2022). URL: https://www.theverge.com/2022/2/15/22935080/facebook-meta-news-feed-renaming-branding-political-content-misinformation.

\textsuperscript{146}Rishi Iyengar. “Here’s how big Facebook’s ad business really is”. In: CNN Business (July 1, 2020). URL: https://www.cnn.com/2020/06/30/tech/facebook-ad-business-boycott/index.html.
updates over minor notifications. Starting in the 2010s Facebook began to leverage the News Feed as their main source of advertising revenue and focused primarily on promoting content that maximized the quantity of “time spent” on the platform, which prompted the introduction of a massive wave of deceptive “clickbait” posts that diverted users’ attention and kept them on the site longer.

While the specific implementation details and the features of these recommendation algorithms differ between each of the prominent social media platforms based upon their structure and the forms of content they prioritize, they all share a similar core framework of relying upon a series of artificial intelligence and machine learning layers to progressively filter and rank posts to potentially display to a specific user. In the case of Facebook’s Feed ranking system, their algorithm first collects an “Inventory” of the full range of posts that a user could be recommended from their network of friends, Groups, and pages that they follow, and the system collects “Signals” on each of these posts that function as the primary datapoints for quantifying popularity and user interaction metrics. Internal reports have indicated that these Signals can comprise “thousands of data points related to the user and post itself” including information about reactions and comments. These can be used to identify posts that the target user will likely interact with or that connect to the user’s recorded interests, summarizing the use of Facebook’s extensive data collection efforts within the operation of the algorithm. Once these Signals have been recorded, the algorithm produces a series of predictions for which posts will be most likely to be relevant to the user before computing a specific “relevance score” for the top options, from which the final rankings are determined to provide the final Feed order.

This general pipeline for producing recommendations is followed by all of the major social media companies for generating the main content flow on their platform, though the specific implementation details typically rely upon a series of machine learning

---

147 Oremus et al., “How Facebook shapes your feed”.
148 Oremus et al., “How Facebook shapes your feed”.
150 Oremus et al., “How Facebook shapes your feed”.
151 Meta Transparency Center, Our approach to ranking.
models to progressively refine the original Inventory into a curated selection of posts. For example, Meta’s Artificial Intelligence division has begun publishing a series of explanations of the algorithmic measures used on their subsidiary platforms and identified 11 separate systems that function in sequence to produce the final recommendations list. This complex tiered framework is comprised of several layers of scoring to rank posts, filtration mechanisms to deprioritize posts that violate community guidelines, and diversity measures designed to prevent a user from solely receiving content from a narrow selection of creators or topics. The computational effect of this series of operations is to produce a single predicted probability that the user will interact or otherwise engage with a specific post (denoted by Facebook as $V_{ijt}$) as a weighted combination of the predicted probabilities $Y_{ijt}$. These probabilities are recomputed on each algorithmic pass and for each individual Signal $X_{ijtc}$, integrating all of these separate features to produce a single composite score for ranking. While this system has been described purely mathematically, this single sophisticated calculation forms the basis for regulating all of the content that is displayed within social media feeds, highlighting the critical role of recommendation algorithms as the driving force behind extensive user engagement with these platforms.

5.6 Pitfalls and Manipulation of Addictive Recommendation Algorithms

Due to the inherent design of recommendation algorithms existing to amplify content that can achieve maximum user engagement, any content creator has the potential to exert a massive impact over the platform’s user population if they can consistently create posts that meet the algorithms’ engagement standards. This is the exact mission of troll farms, bot accounts, and targeted disinformation campaigns, which seek to disseminate

---


153Meta AI, Instagram Feed Ranking System Card.

their misleading or malicious ideas to select populations by manipulating their posts to achieve a higher degree of algorithmic influence. This general strategy has formed the root of the major influence operations described in Chapters 3 and 4, and many of these exploitation techniques are still viable at the present day. In spite of prominent social platforms including Facebook proposing changes throughout their recommendation algorithm frameworks to combat these attacks, the prevalence of disinformation on social media has actually increased since the 2016 election as a result of algorithmic tweaks aimed at promoting increased user interactions.

It is imperative to recognize that while recommendation algorithms are intended to provide a filter on the existing content of a platform they ultimately have a major influence over the types of posts that are submitted to that site and can effectively drive trends in negative or malicious content seeking to exploit the algorithm and reach a large audience. This trend is particularly evident within Facebook’s abrupt shift in 2018 to move from site usage being the primary metric for recommendation to relying upon an internal measure known as “meaningful social interactions” (MSI) that sought to prioritize content with a high number of reactions or that “generate[s] conversation between people.”

Although this change generally had the desired effect of reducing the spread of low-engagement content such as clickbait, it also had the unintentional consequence of immediately presenting a bias towards inciteful content that prompted extensive user discussions. The effects of the MSI system were particularly damaging within the political domain through the idea of influence peddling, where internal investigations concluded that “engagement on positive and policy posts has been severely reduced” by forcing political figures and supporters to become “increasingly reliant on inflammatory posts and direct attacks on their competitors.”

One of the most critical aspects of the MSI system was that its adoption was not driven by development team efforts or internal research but rather a concerted push by CEO Mark Zuckerberg to redefine the structure of the platform’s News Feed. While

---

155 Mosseri, *Bringing People Closer Together*.
5.6. Pitfalls and Manipulation of Addictive Recommendation Algorithms

Zuckerberg contended that the goals of the new algorithmic approach were to “strengthen bonds between users and to improve their well-being” this change was found to backfire almost immediately, with internal reports classifying MSI as an “increasing liability” and asserting that “misinformation, toxicity, and violent content are inordinately prevalent among reshares.”

In spite of these assessments Zuckerberg continued to defend the use of MSI as Facebook’s core new standard for engagement, including directly rejecting many of the proposals from the Integrity Team to reduce the commonality of hate speech or coronavirus vaccine disinformation out of the concern that they could have a “material tradeoff with MSI impact.”

While Facebook’s policies for dictating the direction of the MSI algorithms were evidently flawed, the responsibility for the damages caused by the system also rested upon some of the specific implementation details that prioritized anger and dissent over forms of positive user interaction. In particular, a core feature of the MSI ranking scheme was a reliance upon the options Facebook introduced in 2016 for reacting to a post or comment with different emojis (“‘love,’ ‘haha,’ ‘wow,’ ‘sad’ and ‘angry’”), with MSI ultimately weighting these reactions as “five times more valuable than ‘likes’,” which had served as Facebook’s original means for reacting to content.

Although several engineers within the company had expressed fears that the creation of this system could lead to a rise in the promotion of “controversial” posts that generated heightened emotional responses, it took over two years for Facebook data scientists to formally conclude that angry reactions were “much more frequent” in instances of harmful posts including “civic low quality news, civic misinfo, civic toxicity, health misinfo, and health antivax content.”

This effect was exacerbated even further in the case of divisive content that sparked intense discussion between users replying to the post, with contributions such as a “significant comment, message, reshare or RSVP”...

---

157 Hagey and Horwitz, “Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead.”
158 Dwoskin, Newmyer, and Mahtani, “The case against Mark Zuckerberg: Insiders say Facebook’s CEO chose growth over safety”.
160 Merrill and Oremus, “Five points for anger, one for a ‘like’: How Facebook’s formula fostered rage and misinformation”.

being given the value of 30 likes by the ranking algorithm. While Facebook attempted to counteract this effect and introduce some means of mitigation through “demotions” that divided a post’s composite recommendation score by a constant factor, the lack of a fixed ceiling on these score values and the potential for the exponential spread of contentious posts resulted in these demotions having little effect on the prevalence of negative content across the platform.

Although these vital flaws within Facebook’s algorithms have received the most widespread attention and detailed analysis as a result of the Facebook Files leaks, many of these deficiencies are shared across competing social media platforms and are in fact endemic to the nature of recommendation algorithms. A particularly noteworthy example of this trend occurred within Twitter’s 2021 ML Ethics, Transparency, and Accountability internal study regarding the impacts of their site’s machine learning recommendation tools within their algorithmic Home Timeline; due to the company’s preservation of a chronological timeline of most recent Tweets as an alternative content view, their researchers were able to perform a direct quantitative comparison of the promotion of different political views between the two timelines. Their findings revealed a “remarkably consistent trend” whereby “the mainstream political right enjoys higher algorithmic amplification than the mainstream political left” across 6 of the 7 nations that they performed the experiment in, and furthermore an examination of the spread of U.S. news posts revealed that “algorithmic amplification favours right-leaning news sources.”

While the Twitter ML team did not provide any hypotheses to explain the root causes of this disparity, the researchers did emphasize that the presence of any consistent trends of inequity between political views on their platform is a clear indication of a biased

---

161 Hagey and Horwitz, “Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead.”
162 Merrill and Oremus, “Five points for anger, one for a ‘like’: How Facebook’s formula fostered rage and misinformation”.
recommendation algorithm. By computing measures of both “individual amplification” and “group amplification” for single accounts and aggregated political parties respectively, they were able to directly report the increase in likelihood or “amplification ratio” for an individual to encounter a specific category of post through the algorithmic timeline when compared to the chronological timeline.\textsuperscript{165} In addition to the troubling discrepancies between the amplification ratios for different political parties, the team also identified a more general trend in the amplification of all political content regardless of affiliation (see diagram below).\textsuperscript{166} This consistent promotion of political material reflects the underlying limitation within the structure of recommendation algorithms of trending towards the amplification of any content that yields higher engagement regardless of its impact on society, and in the process provides an implicit argument for the universal use of chronological feeds rather than algorithmic sorting to completely eliminate this potential for abuse.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure5.1.png}
\caption{Graph of group amplification of Tweets from officials in different political parties across each of the seven studied nations from Twitter’s internal study.}
\end{figure}

\textsuperscript{165}Huszár et al., \textit{Algorithmic Amplification of Politics on Twitter}, p. 3.
\textsuperscript{166}Huszár et al., \textit{Algorithmic Amplification of Politics on Twitter}, p. 4.
Chapter 6

Technical Mitigations for Algorithmic Challenges

6.1 Tactics for Improving Moderation and Recommendation Algorithms

While the discussion throughout Chapter 5 served to identify the primary technical flaws within the existing moderation and recommendation algorithms used throughout mainstream social media platforms, the development of these technologies is still a novel and fast-evolving field with significant potential for improvement. Corporations such as Facebook and Twitter continue to implement a wide range of modifications in this domain with the goal of improving user safety and providing more robust responses to disinformation campaigns, but the efficacy of these interventions has been inconsistent. Before addressing the areas where these systems could be enhanced it is necessary to examine the mitigation efforts that these companies have enacted and consider areas where these measures could be expanded upon or developed further to have a stronger effect. However, it is important to note that many of the ideas that have been suggested for handling these challenges have not been put into practice due to executive decisions or financial motives, and as a result the optimal technical strategy may be incompatible with the established business models of these firms.\footnote{Horwitz and Seetharaman, “Facebook Executives Shut Down Efforts to Make the Site Less Divisive.”}
The spectrum of approaches to confronting these threats can be generally divided into two broad categories, with the first consisting of minor additions to existing algorithms that could provide immediate benefits, while the second group is composed of more substantial systematic reforms within the core structure of moderation and recommendation technologies. The vast majority of the tweaks that social media companies make to their sites fall within the former class, with the supplemental nature of these alterations allowing for ad hoc solutions to be quickly tested and deployed even as a crisis is unfolding.\footnote{Timberg, Dwoskin, and Albergotti, “Inside Facebook, Jan. 6 violence fueled anger, regret over missed warning signs.”} However, the relative simplicity of these concepts can also serve as a detriment, as the black box nature of machine learning models can result in a seemingly inconsequential adjustment actually having a sweeping impact on the platform and its users. In contrast to these more straightforward proposals, the latter category would demand the implementation of new technologies aimed at redefining the framework of traditional moderation or recommendation systems in order to address their inherent tendency to promote content that is divisive, inflammatory, or drives user engagement by fostering echo chambers. While such measures would likely be the most effective at limiting the negative influence of social media on society, they must be weighed against the feasibility of convincing these corporations to adopt these more drastic initiatives.

As a result of the consistent pattern of companies such as Facebook rejecting promising algorithmic results based on economic considerations, the majority of the approaches discussed within this chapter are intended to have minimal financial consequences so as to provide a feasible plan of action for improving social media. Therefore, these changes will predominantly be focused on the task of monitoring the quality of user-submitted posts to limit the visibility and spread of harmful content such as disinformation and hate speech. The problem of regulating the contents and influential effects of targeted advertisements falls outside of the scope of this analysis due to the inextricable position of digital ads as the primary revenue stream for social media firms. Furthermore, reforms in the advertising domain would likely require federal policy changes that institute protections against predatory targeting behavior and the excessive collection of user...
The analysis in this chapter is divided into three parts, beginning with Section 6.2 which focuses on mitigation strategies that have already been enacted or tested on current platforms, helping to outline the present state of development in this field. This discussion shifts into the area of more comprehensive algorithmic revisions in Section 6.3, which centers on a series of proposals by former Facebook lead data scientist Jeff Allen designed to combat the spread of disinformation and fake news. Finally, in Section 6.4 I propose an original approach to mitigating the negative effects of algorithmic amplification by treating the outputs of the recommendation pipeline as a direct indicator of the legitimacy of the associated content, taking advantage of the sophistication of existing recommendation techniques to provide a better metric for filtering.

6.2 Building Upon Existing Mitigation Strategies

The development process for moderation and recommendation algorithms on social media effectively functions as a constant feedback loop, wherein company researchers analyze the results of the current implementations and institute updates to adjust algorithmic outcomes. This procedure is epitomized by Facebook’s “remove, reduce, and inform” strategy for governing the handling of harmful posts, with the company focusing on evaluations and results from “three parties: people, publishers, and our community” in order to direct their future efforts. Ultimately, this approach stands as the most viable methodology for developing incremental improvements due to the inability to evaluate new algorithmic features outside of public deployment, as it is nearly impossible to simulate the breadth of content that these algorithms will need to process within any closed internal test setting. As a result, the only practical means to discern how a technical change will impact the propagation of different posts is through live trials with the site’s main user base, which carries significant risks due to the potential for an unproven feature causing users to be exposed to malicious content.

---

In Facebook’s early years the company engaged in these kinds of development efforts with almost no oversight, raising clear ethical issues regarding the legitimacy of participating in these forms of research without user consent and supported only by vague protections in their terms of service and “data-use policy, which governs exactly how it may use users’ data.”¹⁷⁰ The most troubling example of Facebook’s unchecked experimentation occurred during a 2012 study involving almost 700,000 users testing whether their emotions could be influenced by differing recommendations on the site’s News Feed, and the researchers concluded that “emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness.”¹⁷¹ The backlash over the public disclosure of this work from the Facebook Data Science team in 2014 forced the company to replicate the structure of Institutional Review Boards designed to “assess the ethics of human-subject research at academic institutions” and found their own internal IRB, establishing lasting oversight capabilities for all future experimental changes on the platform.¹⁷²

While all the modifications that have been implemented within Facebook in recent years have taken place under the supervision of this IRB, company developers have still been provided with a significant amount of leeway for prototyping new machine learning models or adding new features to refine the results of the News Feed. Several key examples of this precedent have been mentioned in prior chapters, including the “break the glass” measures aimed at protecting the integrity of the 2020 U.S. election (Section 4.3), the heightened reliance of Facebook and Twitter upon AI moderation during the outbreak of COVID-19 (Section 5.4), and the implementation of the Meaningful Social Interactions metric (Section 5.6).¹⁷³ Among these previously discussed instances of

¹⁷³Mac and Frenkel, “Internal Alarm, Public Shrugs: Facebook’s Employees Dissect Its Election Role”.

6.2. Building Upon Existing Mitigation Strategies
algorithmic revision, the “break the glass” tactics stand as the most promising due to their demonstrated success in reducing the spread of disinformation in the days immediately leading up to the 2020 election, yet the company chose to treat these technical features as a temporary intervention strategy only to be used in emergencies or times of heightened tension.\footnote{Timberg, Dwoskin, and Albergotti, “Inside Facebook, Jan. 6 violence fueled anger, regret over missed warning signs”.} In cases such as this where a mitigation program has yielded promising results but is not economically viable for general use, a better strategy would be to continue testing more lenient variations of this approach to identify an optimal middle ground rather than completely abandoning these projects and returning to ineffective traditional algorithms.

Moving outside of the realm of the algorithmic changes discussed earlier in this thesis, the majority of contemporary developments are driven by dedicated bodies including Facebook’s Responsible AI group and Twitter’s ML Ethics, Transparency, and Accountability team, which both work to advance the safety and stability of their respective platforms. For example, the Facebook group has had a primary focus since 2018 of addressing AI bias within their moderation and recommendation systems to prevent these algorithms from “inferring ‘protected’ attributes like race from users’ data” and yielding prejudiced results.\footnote{Karen Hao. “How Facebook got addicted to spreading misinformation”. In: MIT Technology Review (Mar. 11, 2021). URL: \url{https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/}.} Their work on this challenge ultimately produced the Fairness Flow library, which allows for the site’s data scientists and engineers to monitor the equity of newly implemented machine learning models and analyze whether they have any discriminatory consequences. However, the use of this system is completely optional and “none of the teams that work directly on Facebook’s news feed, ad service, or other products” have any requirements to integrate it into their testing frameworks.\footnote{Hao, “How Facebook got addicted to spreading misinformation”.
}

This situation demonstrates the tendency for the most advanced and effective algorithms to not be used in practice, highlighting a course of action for social media platforms to considerably improve their responses to handling detrimental conduct by simply committing to employ the best technical tools available to them.
6.3 Media Literacy and Authoritative Content Measures

In the sparing cases where these cutting-edge technologies are actually deployed, they typically have a notable effect on ameliorating some of the threats faced by social media platforms which reinforces the need for further funding and resources in this area to maximize the capabilities of these companies’ engineering teams. A recent illustration of this is the creation of the “Deep Entity Classification (DEC)” paradigm by a team of Facebook researchers working in conjunction with data scientists at the University of Chicago and Georgia Institute of Technology, collaborating to design a revolutionary theoretical approach for detecting “abusive accounts.”\(^{177}\) While the task of identifying account-based misconduct does not directly address any issues such as disinformation or hate speech, a large proportion of these destructive behaviors are carried out by fake or compromised accounts as a means to protect anonymity or achieve a greater reach. Facebook’s subsequent implementation of Deep Entity Classification resulted in an immediate “estimated reduction of abusive accounts on the network by 27%,” but actually had a far wider effect on limiting the spread of malicious posts.\(^{178}\) As with the prior examples discussed throughout this section, the DEC case calls attention to the opportunity for social media corporations to reform their algorithmic systems purely through the adoption of in-house technologies, which have consistently demonstrated an unrealized potential for enhancing the entire moderation and recommendation pipeline and diminishing the sharing of harmful content.

6.3 Media Literacy and Authoritative Content Measures

Although the proposals covered in Section 5.2 present a promising collection of mitigation strategies for attending to some of the most pressing concerns for modern social media platforms, the selection of algorithmic changes that have already been implemented by these companies represents just a small fraction of the full range of possible designs. The


solutions that fall into this category can take the form of either a theoretical approach conceptualized by an independent external source, or by an internal blueprint created by current or former social media engineers as a recommendation for how to improve their platform. While both perspectives have generated valuable suggestions for reforming social media, the latter group typically provides more concrete and actionable plans due to the familiarity of the designers with the technical specifics and limitations of their site. By appealing to these areas of expertise and integrating more perspectives from developers and data scientists that have exhaustive firsthand knowledge of the underlying nature of these platforms, it is possible to construct a much more informed system for responding to threats than would otherwise be possible under the sole direction of corporate executives.

In internal investigations, one critical area that has received significant attention is Facebook’s persistent promotion of content from troll farms, which function as one of the largest drivers of harmful content on their site. Despite the company’s professed commitment to bolster its defenses against disinformation in the wake of the 2016 election, these changes had little effect on the popularity of troll farm posts, which “reach[ed] 140 million US users per month—75% of whom had never followed any of the pages” in the period leading up to the 2020 election.\(^{179}\) The fact that the majority of users receiving this content had no connection to any of the creators emphasizes the monumental scale of algorithmic amplification occurring under the meaningful social interactions (MSI) paradigm. In addition to the causes of algorithmic amplification discussed in Section 5.6, the disclosures of this troll farm activity also highlighted the defect within Facebook’s recommendation systems of not “penaliz[ing] pages for posting completely unoriginal content,” which allows for malicious actors to directly plagiarize previously successful posts and benefit from their virality to reach a massive user base.\(^{180}\)

This exploitative ability was highly efficacious for the troll farm managers who targeted


\(^{180}\)Hao, “Troll farms reached 140 million Americans a month on Facebook before 2020 election, internal report shows”.
“the same demographic groups singled out by the Kremlin-backed Internet Research Agency (IRA) during the 2016 election,” and they were ultimately able to control “all 15 of the top pages targeting Christian Americans, 10 of the top 15 Facebook pages targeting Black Americans, and four of the top 12 Facebook pages targeting Native Americans.”

Even as Facebook claimed that the interventions described in Section 4.3 were successful in protecting the integrity of the 2020 election, their failure to address the prevalence of troll farms stood out as an example of clear negligence.

While Facebook did not make any deliberate corporate moves to resolve this issue, the pervasive troll farm threat had been identified by internal research dating back to over a year before the 2020 election. Following a year and a half research effort analyzing the propagation of posts from troll farms, Facebook data scientist Jeff Allen published his findings and resigned from the company to launch the Integrity Institute, an independent oversight bureau aiming to "advise policy makers, regulators and the media about how social media platforms work." In Allen’s report, he offers a nuanced examination of the troll farm activity detected throughout the platform and attempts to construct a comprehensive overview of the ramifications of this trend, directly implicating Facebook in its complicity in "empowering inauthentic actors to accumulate huge followings for largely unknown purposes." Building upon this foundation, Allen provides a detailed statistical analysis of the mechanisms by which Facebook’s engagement-based MSI recommendation rankings can be manipulated, revealing that “content that scores highest in these fundamental models heavily skews towards content we know to be bad.” Furthermore, the site’s lack of detection mechanisms for duplicated content ensures that any copied submission “will likely get a similar score the second time it is posted,” allowing for blatant exploitation techniques to be employed with no consequences.

---

181 Hao, “Troll farms reached 140 million Americans a month on Facebook before 2020 election, internal report shows”.
184 Allen, How Communities Are Exploited On Our Platforms, p. 10.
In order to grapple with these massive challenges, Allen proposes a possible framework based upon Google’s approach to maintaining the quality of their search results and preventing malicious sites from taking advantage of the algorithms to maximize their search ranking. The inherent structure of Google’s search logic is actually responsible for their strong defenses against abuse, with their PageRank algorithm providing a “Graph Authority” score that quantifies a site’s legitimacy based upon the reliability of the sites that link to it. Allen divulges that Facebook actually possesses their “own implementation of a graph based authority measure” that allows for the scoring of individual pages, and he notes that within internal tests “90% of the Troll Farm Pages have exactly 0 Graph Authority” so the simple integration of this metric into recommendation algorithms could immediately eliminate the majority of troll farm content. By extending the scope of the filtering process for removing harmful content from a user’s “Inventory,” Facebook would effectively be able to institute their own methodology for objective content creator quality rankings that could prevent malicious users from ever establishing a significant reach on the platform. If this initial Graph Authority integration proved to be successful, these quantitative measures of validity could ultimately be expanded into areas such as approving advertising partnerships or regulating the results of the site’s Suggested Groups feature, serving to protect all branches of the Facebook network from exploitation by illegitimate parties.


6.4 Recommendation Scoring as an Inverted Indicator of Legitimacy

A consistent element present throughout the frameworks discussed so far in this chapter is that they have all been grounded within the traditional algorithmic architecture used across mainstream social media platforms in which moderation functions as a discrete stage within the overall recommendation pipeline. As described in Section 5.5, the primary purpose of moderation machine learning systems within modern platforms is to function as the initial filtration step for removing posts from consideration for being recommended, in effect relegating moderation to being a subsidiary concern beneath the priority of delivering engaging content to users. This structure is illustrated through the format of Instagram’s primary Feed rankings, in which the single initial pass through a “Filtering System” is followed by 9 successive tiers of ranking and sorting models to yield the final ordering of a user’s Feed, demonstrating the clear imbalance in priority
between the treatment of moderation and recommendation.\textsuperscript{187}

The significantly higher complexity of these recommendation layers in comparison to the solitary moderation process ensures that the former area will be allocated superior resources and larger teams of dedicated developers, limiting the ability for the moderation technologies to be substantially improved. Furthermore, the pre-filtration approach used by Instagram also applies pressure to the moderation staff to maintain lenient standards so as to avoid filtering out posts that could achieve large recommendation scores, effectively introducing an implicit bias into the process of pruning content. Unless a post displays an incontrovertible violation of one of the site’s community standards, the overall goal of stimulating user engagement incentivizes the moderation layer to allow all borderline posts to pass and be ranked naturally by the recommendation steps.

Although moderation and recommendation are typically depicted as two distinct algorithmic problems, the breadth and comprehensiveness of the numerous factors integrated into modern recommendation systems causes these models to implicitly encode indications of whether a post was created by a genuine source. This phenomenon is directly rooted in the extreme degree of complexity of these current pipelines, as demonstrated by the admission of a Facebook integrity official that their ranking algorithm takes in “more than 10,000 different signals to make its predictions about a user’s likelihood of engaging with a single post.”\textsuperscript{188} Facebook’s systems integrate such a wide range of information about the underlying features of each individual piece of content that an item’s recommendation score can be seen to provide some indirect measure of legitimacy, with posts from malicious users such as troll farms frequently having excessively high-ranking values. This was actually corroborated within Jeff Allen’s research report in a provisional stage of analysis that compared the ”value per view” obtained in advertising revenue against a post’s meaningful social interactions (MSI) score for both troll farms and verified partner pages.\textsuperscript{189} Allen discovered a perceptible statistically significant difference between these metrics for the two page categories, supporting the conclusion

\textsuperscript{187}Meta AI, \textit{Instagram Feed Ranking System Card}.
\textsuperscript{188}Oremus et al., “How Facebook shapes your feed”.
\textsuperscript{189}Allen, \textit{How Communities Are Exploited On Our Platforms}, p. 11.
that the outcome of a recommendation ranking system such as MSI can serve as a proxy for modeling the authenticity of various content creators.

![Graph from Jeff Allen’s paper spotlighting the difference in the distribution of MSI scores between troll farms and partnered pages.](image)

**Figure 6.2:** Graph from Jeff Allen’s paper spotlighting the difference in the distribution of MSI scores between troll farms and partnered pages.

While these preliminary results provide promising signs for the existence of a correlation between recommendation score and legitimacy, the trends exhibited by Jeff Allen exist only at a platform-wide level across a wide selection of pages and do not provide any mechanism for assessing the quality of individual posts. However, further research efforts in this field have scrutinized this effect at a lower level and have established compelling evidence for the claim that there are in fact diverging patterns in the propagation of individual posts based upon their authenticity. A groundbreaking 2018 study by a group of MIT media scientists comparing the dissemination of fake and genuine news stories on Twitter found that “falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information, and the effects were more pronounced for false political news than for false news about terrorism, natural disasters, science, urban legends, or financial information.”

---

research team found that fake news stories typically spread at a rate and scale of 1-2 orders of magnitude greater than analogous factual reports, and this relationship was still maintained even after “remov[ing] all of the bots in our dataset” and analyzing just the subset of human-run accounts.\footnote{Peter Dizikes. “Study: On Twitter, false news travels faster than true stories”. In: MIT News (Mar. 8, 2018). URL: \url{https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308}.}

Based upon the existence of these predispositions within the foundations of social media content propagation, it is possible to conceptualize a revolutionary approach to filtering that treats recommendation scores as a primary metric for judging the legitimacy of posts rather than serving as the final output in the algorithmic chain. This new structure would fundamentally redefine the traditional linear recommendation pipeline into a recursive loop pattern that feeds the ranking output from the first pass of the algorithm back into the moderation layer, which can assess this score in conjunction with other relevant signals in order to ascertain whether a given post matches the typical profile of content from a problematic source such as a troll farm or disinformation campaign.

Due to the fact that the recommendation score has a direct correspondence to the rates of promotion and spread across the platform, any post with an exceptionally high ranking that falls into a sensitive category such as political speech or public health information will have a significantly higher risk of containing harmful content. As a result, the use of recommendation scores as an “inverted indicator of legitimacy” enables social media platforms to have a far more robust and pragmatic response to handling cases of borderline content that would not be filtered out by current moderation algorithms, with this novel solution instead applying demotion procedures to these posts to prevent their continued spread. Ultimately, the expanded reliance upon existing recommendation systems under this theoretical solution would allow for social media companies to better leverage their most sophisticated technical tools for achieving more thorough moderation and providing safer platforms that protect both their user bases and society at large from the threats of malicious actors.
Chapter 7

Conclusions and Policy Proposals

7.1 Lacking Self-Regulation and the Necessity of Policy Reforms

Although the technical ideas offered throughout Chapter 6 present an auspicious framework for the algorithmic refinement of social media platforms to better protect against threats such as disinformation and hate speech, these proposals are insufficient on their own for inducing sweeping reforms within the social media industry. As noted in Section 6.1, despite the promising results of these suggested interventions and the opportunity to greatly improve the safety of these sites, corporate executives have typically been reluctant to embrace efforts to institute stronger moderation or content filtering systems. This opposition has been driven historically by the dependence of leading platforms on digital advertisements for generating virtually all of their revenue, forcing these companies into a business model that encourages the maximal usage of the platform.\(^{192}\) In this uncompromising quest for profit, the attention of users is effectively commodified such that their engagement with the site becomes what Mark Zuckerberg refers to as a “north star,” driving the entire development of Facebook and its affiliated services to function in support of this objective.\(^{193}\) Unfortunately, this financial motive comes directly into conflict with the stated mission of confronting the societal challenges and risks to the public good that are posed by social media, and through a combination of corporate greed and economic responsibility to shareholders these companies have

\(^{192}\) Iyengar, “Here’s how big Facebook’s ad business really is”.

\(^{193}\) Dwoskin, Newmyer, and Mahtani, “The case against Mark Zuckerberg: Insiders say Facebook’s CEO chose growth over safety”.

consistently prioritized the optimization of engagement metrics over the implementation of more aggressive measures to regulate user-submitted content.

As a result of this ongoing tendency to favor leniency and user engagement over safety measures, it appears necessary for there to exist some form of outside influence in order to compel social media companies to adopt stricter approaches to content management. In some cases, the obligation to better regulate posts has arisen from revelations about the dangers of social media, which was summarized in Facebook’s response to the Stop the Steal movement after the 2020 election. As described in Section 4.4, the company only moved to institute extensive measures to halt the formation of Stop the Steal groups due to the massive attention generated by the January 6th Capitol riots, undermining the claim that they would responsibly handle threats in the absence of pressure.194 In addition to influence from the general public this pattern has also been replicated within private supervisory groups, as was illustrated in Section 4.2 with Facebook’s own Oversight Panel chastising the company for their concealment of the XCheck program and demanding greater transparency.195

While these examples have all resulted from the emergence of public disclosures of the internal activities of social media firms, an additional pathway for instigating reforms could be employing legislative measures as the most comprehensive possible mechanism for mandating compliance. In some cases, such as the 2016 U.S. Presidential election, this political pressure has not actually taken the form of any legislative efforts but rather has relied upon the use of Congressional hearings to publicize the extent of the damaging features of social media and effectively coerce platforms into revising their algorithmic strategies to avoid further exposure and investigation. However, while U.S. legislators have frequently raised the issue of regulating social media platforms and have discussed a range of different approaches to tackling this challenge, bipartisan support for regulation has been sparse and little agreement has been reached over the optimal

194 Timberg, Dwoskin, and Albergotti, “Inside Facebook, Jan. 6 violence fueled anger, regret over missed warning signs”.
195 Oversight Board, Oversight Board demands more transparency from Facebook.
course of action.\textsuperscript{196}

\section*{7.2 Ineffective Attempts at Legislating Social Media}

The most prominent proposals for legislating the reform of American social media platforms have included restructuring or eliminating Section 230 of the Communications Decency Act so as to “hold the online platforms accountable for profits they earn from deliberately disseminating harmful content,” installing direct legal regulations on moderation and recommendation algorithms to limit the ability for companies to tweak their implementations for the sake of profit, and pursuing anti-trust legislation against Facebook and the other major social media companies to reduce their massive influence over the entire industry and allow for fairer competition.\textsuperscript{197} While the U.S. Congress has failed to successfully enact any of these general types of legislation and has largely allowed for the American social media industry to evolve unchecked over the past two decades, bills falling into one of these three classes have been drafted frequently in recent years but have lacked the level of consensus and endorsements necessary to yield a viable legislative plan. Despite their continued inefficacy, these broad categories remain the main topics of debate over the optimal methodology for regulating Facebook and other leading platforms, providing a relatively narrow scope of potential avenues for formulating legal responses.

The first area regarding modifications to Section 230 has received particular focus in the past year through the introduction of the Justice Against Malicious Algorithms Act, which would seek to restrict this provision of the Communications Decency Act by waiving its protections in cases where the company “knowingly or recklessly uses an algorithm or other technology to recommend content that materially contributes to


Chapter 7. Conclusions and Policy Proposals

physical or severe emotional injury.” However, the feasibility of actually imposing changes to Section 230 that could be upheld in court is constitutionally dubious due to the strong linkage between the original act and the First Amendment’s protections on the freedom of speech, with any federal limitations on the scope of protected discourse on social media potentially being construed as a constraint on the reach of public speech in the modern sphere. Moreover, as was emphasized in Section 5.2, the existence of the Section 230 law was indispensable to the development of social media, which could not have existed if platforms were instead treated as “publishers, making them directly liable for the content they host.” Rather than engendering a significant boost to the strength of moderation technologies, limitations to Section 230 would subvert the entire order of social media and force companies to greatly scale back on the freedom of speech allowed on their sites, minimizing the appeal of this tactic.

The latter two categories of suggested legal interventions also face similar problems in terms of defensibility against legal appeals from social media corporations and whether they would successfully achieve reforms or would fail in practice to have a positive effect on improving the domain of social media. The second group of recommended laws dealing with direct adjustments to the content of algorithms is beset by similar struggles to the approaches relating to Section 230, with both scenarios posing a threat to the constitutional protection of free speech under the First Amendment. In particular, recent proposals including the 2022 Social Media NUDGE Act that would require platforms to implement specific technical features such as a mandate that users cannot share news sources without first interacting with the linked article seem to present immediate challenges to the free speech test. Although this concept was developed as an innocuous attempt to reduce the spread of false or misleading news by forcing users to actually

---


199 Zorthian, “Washington Wants to Regulate Facebook’s Algorithm. That Might Be Unconstitutional”.


7.2. Ineffective Attempts at Legislating Social Media

Click on the link as an indicator that they engaged with an article beyond just reading its headline, this policy could easily be interpreted as a legislative restriction on speech and fail to comply with the terms of the First Amendment. Furthermore, it is unclear that this system would even function as intended or if users would quickly learn to exploit it by clicking on articles without reading them just for the sake of enabling the permission to post, such that this tool would not actually provide any genuine measure of the user’s awareness of the contents of a shared link.

Finally, the massive scale of modern tech industry giants has sparked frequent calls for these firms to be broken up through the enforcement of federal antitrust legislation and oversight from the federal government’s dedicated agencies for handling corporate misconduct. The most pressing example of this debate concerns the Federal Trade Commission’s ongoing lawsuit against Facebook, with the FTC contending that the company has been “illegally maintaining its personal social networking monopoly through a years-long course of anticompetitive conduct.”\footnote{Federal Trade Commission v. Facebook, Inc. No. 20-3590 (JEB) (D.D.C. June 28, 2021). URL: https://www.ftc.gov/legal-library/browse/cases-proceedings/191-0134-facebook-inc-ftc-v.
}

Under this so-called “buy-or-bury” approach, Facebook leveraged exceedingly aggressive tactics in buying out competing platforms such as Instagram and WhatsApp to ensure their continued dominance over the industry, while “cutting off third parties from its platform” that failed to negotiate acquisition deals.\footnote{Cecilia Kang. “A Facebook antitrust suit can move forward, a judge says, in a win for the F.T.C.”. In: The New York Times (Jan. 11, 2022). URL: https://www.nytimes.com/2022/01/11/technology/facebook-antitrust-ftc.html.}

Although the FTC presented a strong argument corroborating Facebook’s monopolistic tendencies, the ambiguity of existing antitrust laws and the difficulty in establishing the extent of a corporation’s “market power” ultimately hinder their case and undermine the efficacy of this approach.\footnote{Gilad Edelman. “Lina Khan’s Theory of the Facebook Antitrust Case Takes Shape”. In: The Verge (Aug. 19, 2021). URL: https://www.wired.com/story/lina-khan-theory-facebook-antitrust-case-takes-shape/.

Furthermore, it is unclear whether a successful antitrust verdict over Facebook would actually bring about advancements in the safety of social media or whether the resulting competition in the sector would in fact lead to more severe abuses akin to the early development of the industry described in Chapter 2. Overall, the current methods being pursued by the American government
for regulating social media seem to possess significant inherent disadvantages or conflict with the nation’s established legal framework, impeding the potential for effective progress to take place.

7.3 Limiting Revenue Incentives and Compelling Maximum Enforcement

Despite the clear shortcomings within the existing American response to the federal oversight of social media, other nations have outlined more robust and practical approaches for engaging in this type of reform effort. One of the most successful recent examples of this occurred within Australia in February 2021 through the ratification of the News Media and Digital Platforms Mandatory Bargaining Code, which was introduced to “support the sustainability of the Australian news media sector by addressing bargaining power imbalances between digital platforms and Australian news businesses.” This law serves to ameliorate the modern problem of traditional news providers losing the majority of their revenue to social media sites that share their content without providing any financial support for journalism, and the Code decrees that platforms such as Facebook and Google must “pay publishers if they host their content.” While this law serves to revitalize the floundering conventional news industry, it was originally met with intense backlash from the involved tech companies including Facebook, which immediately announced that it would “restrict publishers and people in Australia from sharing or viewing Australian and international news content” in order to avoid having to comply with the financial demands of the Code. After a week of tense debates and public outcry from Australian users who were suddenly unable to access any news


media through the platform, the company ultimately surrendered to this criticism and established a deal to selectively negotiate payments to “support the publishers we choose to.”\textsuperscript{208}

Facebook’s capitulation during the passage of the Australian Mandatory Bargaining Code highlights the potential for effective legislation that incentivizes change within the tech industry, laying out a possible blueprint for the American government to follow in instituting similar measures. Although the requirement that an internet platform must compensate any news providers for the use of their reports appears to serve solely as financial support for journalism, it can have the side effect of reducing the prevalence of fake news and misinformation. Within the new Australian model, Facebook must reach separate agreements with individual news sources in order for their stories to be shareable on the platform, directly inhibiting the ability for unsubstantiated and unverified stories to be posted by Australian pages and users.\textsuperscript{209} The specificity of this law to the Australian user base demonstrates the capacity for American congressional action to demand that social media platforms adhere to similar domestic policies. This case confirms Facebook’s capability to selectively target a specific set of users and moderate their posts with a high degree of precision to ensure compliance with a national law, raising the possibility for comparably strict initiatives to be enacted for the American audience. Although the American legal system differs from that of Australia due to the latter country’s lack of constitutional protections for the freedom of expression, it is critical to recognize that the First Amendment only safeguards against governmental interference and by extension “the right to free speech doesn’t extend to a right to have that speech amplified by algorithms.”\textsuperscript{210}

Following this paradigm for effective reform, the American government should turn their attention towards the formation of an agency that can compel social media companies to leverage their best existing technical tools to support user safety. It is clear


\textsuperscript{209}Morrison, “Why Facebook banned (and then unbanned) news in Australia”.

from the discussion in Section 7.2 that Congress cannot explicitly specify the nature of the algorithms that a platform must implement without violating that company’s freedom of expression by forcing them to include particular features within its source code. However, Congress would be fully warranted in mirroring the Australian precedent by establishing a system of financial incentives that give these platforms an economic obligation to uphold their duty in protecting against threats. The introduction of significant fines that directly penalize companies for instances of misinformation on their site would likely induce them to rely upon more innovative solutions. These sanctions could be intensified through the application of the Sarbanes-Oxley Act of 2002 which allows for company executives to be held fiscally responsible for their corporation’s transparency and could render “Mark Zuckerberg and the Facebook board liable for failures to implement their terms of service.” The enforcement of this punitive scheme would likely be too complex for Congress to handle directly and could merit the foundation of a new executive Social Media Oversight Commission akin to the Federal Trade Commission or Securities and Exchange Commission that is specifically designed to assess the actions of social media corporations.

The formation of this proposed oversight body could serve as the final necessary component for government regulation to maintain accountability throughout the industry and persuade these companies to adopt some of the more aggressive technical measures put forward in Chapter 6. The entire history of the abuses carried out by social media platforms has occurred in the absence of any form of consistent federal supervision; consequently, the establishment of an agency dedicated specifically to this field would usher in a new era of social networks that are forced to prioritize the public good over their raw sources of revenue. The application of political pressure could pave the path for these companies to implement more robust and systemic algorithmic revisions such as the Graph Authority metric analyzed in Section 6.3 or the “invented indicator of legitimacy” approach proposed in Section 6.4. These stronger reforms provide the most comprehensive methods for mitigating the presence of malicious content and allowing

211Cusumano, Yoffe, and Gawer, “Pushing Social Media Platforms to Self-Regulate”.
the company to pass oversight standards, and as a result the conjunction of these technical and policy components could revolutionize the treatment of harmful conduct on social media. Only through this process could any social media platform ever truly fulfill Facebook’s mission statement of “bring[ing] the world closer together,” maintaining the power to connect the global community while eliminating the most deleterious impacts on society.\textsuperscript{212}

Bibliography


BBC News. “Facebook admits it was used to ‘incite offline violence’ in Myanmar”. In: *BBC* (Nov. 6, 2018). URL: https://www.bbc.com/news/world-asia-46105934.


Bobrowsy, Meghan and Ryan Tracy. “Facebook Oversight Board Launches Review of Company’s XCheck System”. In: *The Wall Street Journal* (Sept. 21, 2021). URL: https:


“Facebook’s long and halting fight against misinformation”. In: AP News (Oct. 18, 2020).
URL: https://apnews.com/article/virus-outbreak-donald-trump-conspiracy-theories-misinformation-elections-36cca2a5fb8911132b0d8f0b6a3e5a31.


— “Russian propaganda may have been shared hundreds of millions of times, new research says”. In: The Washington Post (Oct. 5, 2017). URL: https://www.washingtonpost.com/news/the-switch/wp/2017/10/05/russian-propaganda-may-have-been-shared-hundreds-of-millions-of-times-new-research-says/.

Timberg, Craig, Elizabeth Dwoskin, and Reed Albergotti. “Inside Facebook, Jan. 6 violence fueled anger, regret over missed warning signs”. In: The Washington Post (Oct. 22,


Wilson, Steven Lloyd and Charles Wiysonge. “Social media and vaccine hesitancy”. In: BMJ Global Health 5.10 (2020). DOI: 10.1136/bmjgh-2020-004206. eprint: https://gh.bmj.com/content/5/10/e004206.full.pdf. URL: https://gh.bmj.com/content/5/10/e004206.


Bibliography


