THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF INFORMATION SCIENCES AND TECHNOLOGY

ANALYZING SOCIAL MEDIA PLATFORMS TO ESTABLISH EVALUATION METRICS FOR PROPENSITY OF S.C.A.R.E. ABUSE

AUSTIN THOET
SUMMER 2022

A thesis submitted in partial fulfillment of the requirements for a baccalaureate degree
in Security and Risk Analysis with honors in Security and Risk Analysis

Reviewed and approved* by the following:

Michael Hills
Professor of IST
Thesis Supervisor

Dinghao Wu
Professor of IST
Honors Adviser
* Electronic approvals are on file.

ABSTRACT

The negligent design of social media platforms allows malicious actors to manipulate and exert control over unsuspecting users. This collective of abusers is established as S.C.A.R.E., representing Scammers, Chaos Causers, Advertisers, Radicalizers, and Election Interferers. This group thrives on four common structural weaknesses of social networking platforms, which include anonymity, personalization, engagement priority, and content moderation. The use of evaluation metrics determines the extent each weakness exists on a platform. These calculations grade each social media platform's likelihood of abuse by S.C.A.R.E. actors based on the presence of structural flaws. After Facebook, Twitter, Youtube, and Reddit, are scored, research can be conducted on structural improvements for each of the detailed weaknesses. This begins the path to establishing a modern framework for social media design. This study provides interested parties with a path for development that prevents endangerment of their userbase to mental health issues, loss of decision-making autonomy, stolen property or identity, as well as countless other harmful outcomes. Ultimately, the purpose of this study is to help the average social media user avoid exploitation and participate freely in a digital age.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor, Michael Hills. Without his insight and guidance, this thesis would not have been possible. Additionally, I would like to thank my honors advisor, Dinghao Wu. Finally, I want to thank my family and friends who supported me until the end.

Chapter 1

Introduction

Humanity is facing a crisis of its own creation. Innovation allowed humans to reach further than ever before and connect with one another to an unprecedented extent. Eventually, this progress-driven mindset launched the world into a globalized ecosystem, albeit existing on a digital foundation. It followed that people wanted a way to socialize with one another within this ecosystem. Then, social media delivered on this dream. Absolutely astonished by this novel invention, people flocked to set up an account and join their friends and family. In society's haste to embrace this miracle, no one paid attention to the adverse ramifications or to those seeking to abuse platforms for personal gain. Approximately 20 years later, these consequences manifest themselves in worse ways than ever conceived. Adverse effects range from exacerbating the mental health crisis to increasing levels of crime to threatening democracy to countless others on a seemingly never-ending list. These costs did not become fully evident until real-world events happened as direct responses to online campaigns such as mass shootings, altered election results, suicides, and countless horrendous events. Witnessing this wide array of consequences, it is logical to blame the companies responsible for social media's creation, but should they accept complete liability?

Not one of the social media goliaths examined in this study, Facebook, Twitter, YouTube, or Reddit, rose to prominence on the basis of causing damage to society. Rather, they promised unparalleled possibilities in connection, education, and entertainment[66]. As expressed above, when progress without proper concern given to secure design plagues an invention, it leads to the absolute corruption of the core purposes, and in this case, of mainstream applications. Initially, society wished to keep social media platforms on pedestals as they loved the abilities gained from them. However, everything changed when corporations' direct knowledge of design flaws and their subsequent cover-up was leaked and became well-documented. Focus and public perception shifted from pardoning for negligence to pressing for accountability[6]. This generation is setting a higher standard for the acceptable standards innovators must meet. Yet, it still proves difficult to hold these companies liable as the root causes are far from evident. That is why the structural mechanisms that allow the malicious users to have success must be targeted, isolated, analyzed, and eradicated.

Due to the intentional negligence of social media companies and governments, no effective frameworks exist to combat these issues. While this claim might be viewed as bold and subjectively harsh, its criticality to the online realm's future development proves justified. Thus, this thesis serves to classify these actors who abuse social media into a collective, which will inform an in-depth analysis of how each member operates on social media. Specifically, a concentration of what structural weaknesses allow for them to thrive. Success in connecting social networking flaws as mutually shared within this group exposes a means to measure a platform's proclivity to be exploited. Metrics allow for continuous tracking and adjustment in response to updates and improvements, moreover, progress. Ultimately, this process informs a refined structural framework for social media platform creation. If nothing is done, the next generation of social networking guarantees the same implications. The urgency and necessity of this evaluation framework stem from the current development of the metaverse and Web 3.0. Both promise immense benefits, but the flaws within the existing platforms need patching.

Captology

Throughout history powerful groups employed propaganda and persuasive techniques to steer public opinion and action. Eventually, pioneers refined these mechanisms into highly successful strategies for controlling human beings' opinions. While in some cases this curation proved beneficial for society, the inclination to make money and claim power dominated their use. In the present, this issue worsens every day at a magnitude unprecedented and unpredicted up until 1996. This was when computers started to garner interest, specifically as tools of persuasion. The foremost expert on viewing computers as persuasive technologies, B.J. Fogg a professor at Stanford University, coined the term captology for this utilization[15]. Captology denotes actions taken regarding the "design, research, and analysis of interactive computing products created for the purpose of changing people's attitudes or behaviors."[15] Before this definement the potential of computing as seen today was severely underestimated. Consequently, a few select groups gained absolute dominion over the human population. Primarily through the exploitation of captology principles on social media. It is in this digital age of social media, that we find our minds compelled to make decisions that are not our own, but rather to serve the objectives of these powerful actors. They furthered their ambitions successfully. In doing so subverted the sanctity of thought autonomy through advanced neurochemistry knowledge. A minor price of ethics to pay for the capabilities of a god.

Edward Bernays, "the father of public relations", witnessed this dire situation in 1928 in his book Propaganda. He claimed that "there are invisible rulers who control the destinies of millions."[14] Seemingly outlandish when published, it now proves to be descriptive of the present-day situation of social media; where the rulers are companies and financers, and the destinies are in the billions as global social media users surpassed 4.2 billion in 2021[21]. A chilling foreshadowing of the dilemma we face in the 21st century. These invisible rulers have a proclivity to distort the principles of captology for corrupt purposes. This study assigns the malicious actors as S.C.A.R.E., Scammers, Chaos Causers, Advertisers, Radicalizers, and Election Interferers.

S.C.A.R.E. Qualifications

The acronym is the first outline of social media threat actors and serves to be a baseline for future studies. Building on B.J. Fogg's work pertaining to persuasive technology[15], this study defines an agglomeration of actors who seek to exploit social media users. In-depth analysis and case studies inform the process through which these individuals, and groups, are selected. Therefore, a combination of four qualifying characteristics must be present to gain admittance to this collective: win-lose intention, manipulation of users, growth with social media, and exploitation of social media structure. No exception is made for nation-state entities to preserve the impartiality of judgment.

Win-Lose Intention

It is important to distinguish that if technology is created for one intention but has a negative persuasive outcome on the audience that differs from the initial intention, it does not fall within captology. Instead, it is the alignment of an intention with an outcome that determines whether a computer is wielded as a persuasive technology mechanism[15]. This premise informs the establishment of the first criteria a S.C.A.R.E. actor must meet. Each must act with an intention that directly relates to a negative outcome for users and positive effects for themself. A true embodiment of the win-lose mentality in business, which normally strains relationships. In an online environment, there are endless victims, so a win-lose situation has minimal to no consequences on the initiating side.

Manipulation of Users

If in order to achieve their objective, it becomes necessary to change what a user thinks or believes, a manipulation has occurred. The Cambridge Dictionary defines manipulation as "controlling someone or something to your own advantage, often unfairly or dishonestly"[102]. In this sense, any action taken to abuse captology in an unfair manner that changes a belief or thought process is a manipulation of the user. Contrasting with the first criteria, the manipulation of users focuses directly on the actor taking steps to control and/or change what the user believes. The Win-Lose intention criteria is the motive and the manipulation of user criteria is the action.

Growth with Social Media

The pace and magnitude of social media adoption were the distinguishing characteristics. However, that same pace and magnitude exacerbated the growth of malicious actors' success. Noting this rapid growth of social media, the third criteria of S.C.A.R.E. requires the actor to increase their own growth rate as social media became more popular. This serves to prove causation between the two. A correlation deems an incendiary impact of social media on the success of the actor. Overall social media embracement statistics establish a proper time period to compare.

Social media adoption rates increased from 5% to 72% from 2005 to the onset of 2021 among American adults[3]. This is a 1,440% increase in its userbase size. Furthermore, elderly Americans distort this adoption figure significantly, which disproportionately biases the perception of younger age groups. For instance, the 18-29 and 30-49 year-old categories demonstrate 88% and 78% adoption rates respectively, which is much higher than the combined percentage and increase[64]. These rates are historically matched only by the fastest adoption trends of technologies deemed essential in the United States.

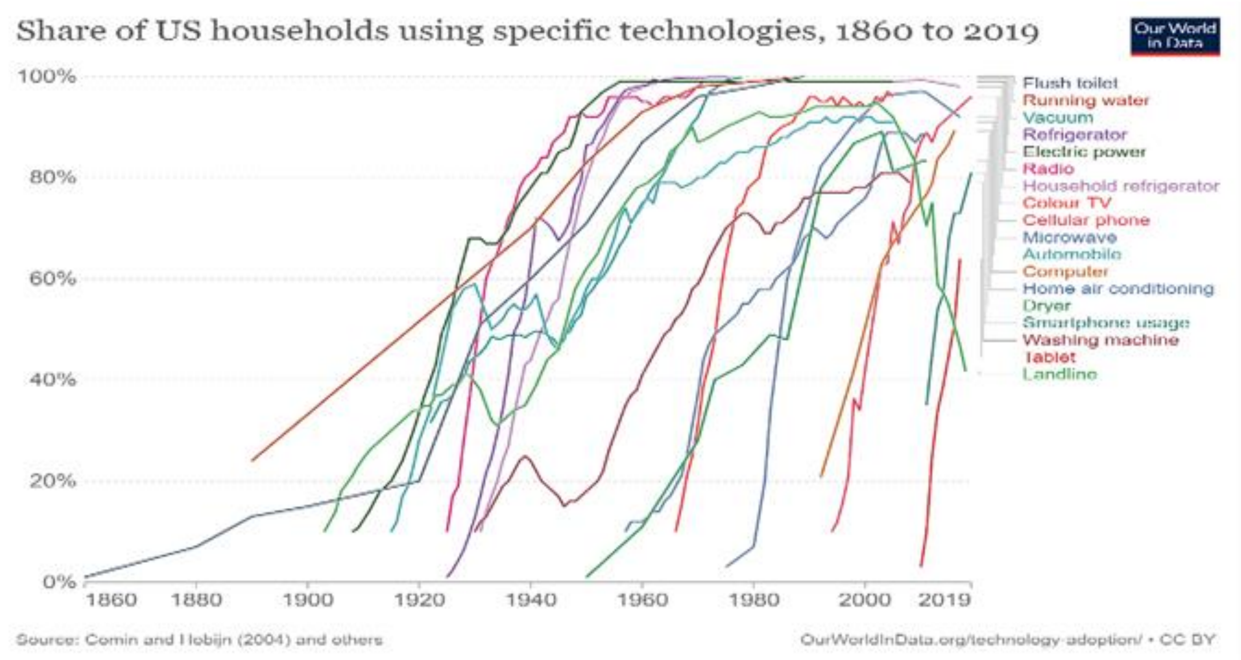


Figure 1. Adoption of essential technologies[4].

The figure above proves that social media was embraced in a shorter time span than household refrigerators[4]. This further supports the classification of social media as a requisite component of living in modern society and a high-growth entity between 2005 and 2021.

Each potential actor must display similar growth levels over this approximate period. In order to measure this growth, every actor requires varying types of data. These types range from growth in the public interest, economic impact, frequency of actor’s abuse, and share of success attributed to online platforms. Any group that meets these criteria based on the data points outlined, can undergo analysis to identify key features that explain the interconnectivity between social media and it.

Exploitation of Structure

The final criteria for admittance into S.C.A.R.E. is the exploitation of the social media structure. After the motive, action, and correlated growth are proven, the exacerbating effect of structural weaknesses is evaluated. If the reason for the actor’s growth is due to flaws and can be stopped through effective security frameworks, admittance is awarded. This factor proves the focal point of the case studies portion and the remaining areas of the thesis.

S.C.A.R.E. Actors

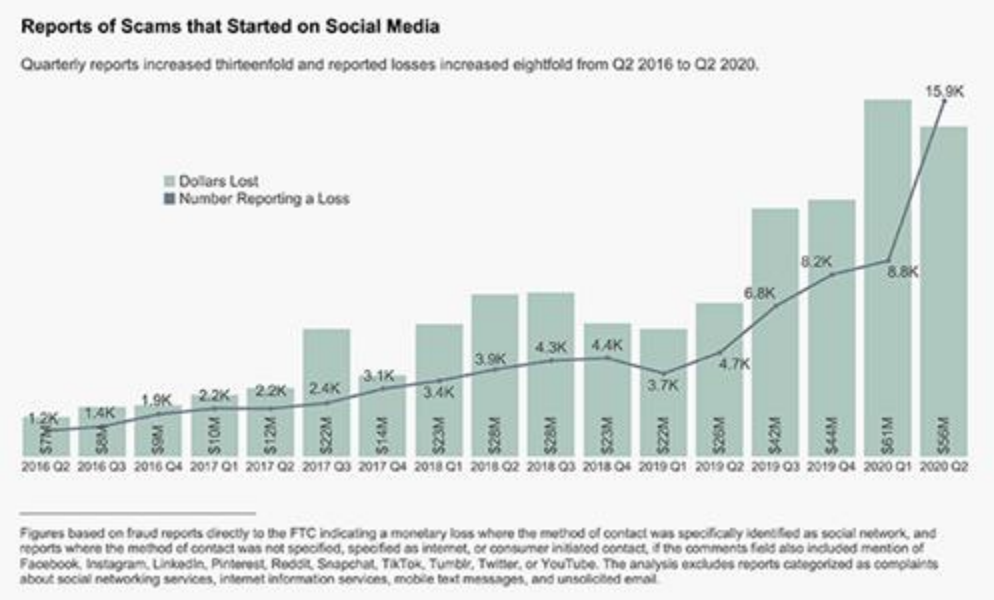
This study chooses the five actors, Scammers, Chaos Causers, Advertisers, Radicalizers, and Election Interferers, due to key characteristics predicated on the author’s active observation of social media during the preceding five years. The first three criteria are analyzed here and fulfillment of all three grants an actor admittance into S.C.A.R.E.. The final criteria, Exploitation of the Structure, will be explored in the Case Studies component. Success in demonstrating each S.C.A.R.E. member exploited social media structure proves the validity of the S.C.A.R.E. framework as a system.

Scammers

Cambridge defines a scammer as “someone who makes money using illegal methods, especially by tricking people.” [24] At the most basic level, a scammer convinces an individual to give them money, information, or another possession under a false pretense and consequently to their detriment. This embodies the Win-Lose criteria as the scammer can only gain when their target loses.

The types of scams are limited only by imagination but the primary tenets traditionally rely on deceptive practices to achieve financial or informational gain. Furthermore, the majority of scams on social media are not based on novel methods but rather bear resemblance to real-world fraud schemes[26]. These schemes pry on human nature and exploit inherent psychological patterns to change the way a person thinks. Through this, the criteria of manipulation of the user is satisfied.

The advent of social commerce was a shopping revolution and was considered an established practice by 2012. This rapid acceptance and the spread of social media led to the online scamming industry. Growth exceeded expectations due to early narratives that social commerce was “media hype or a business fad.”[25] Consequently, this created a prime environment for scammers to operate. These existing schemes’ improved success, when migrated to social media platforms, can be shown by the rate of total loss expansion reported to the Federal Trade Commission.


Reports of Scams that Started on Social Media

Quarterly reports increased thirteenfold and reported losses increased eightfold from Q2 2016 to Q2 2020.

Figures based on fraud reports directly to the FTC indicating a monetary loss where the method of contact was specifically identified as social network, and reports where the method of contact was not specified, specified as internet, or consumer initiated contact, if the comments field also included mention of Facebook, Instagram, LinkedIn, Pinterest, Reddit, Snapchat, TikTok, Tumblr, Twitter, or YouTube. The analysis excludes reports categorized as complaints about social networking services, internet information services, mobile text messages, and unsolicited email.

Figure 2. Scams on Social Media[23].

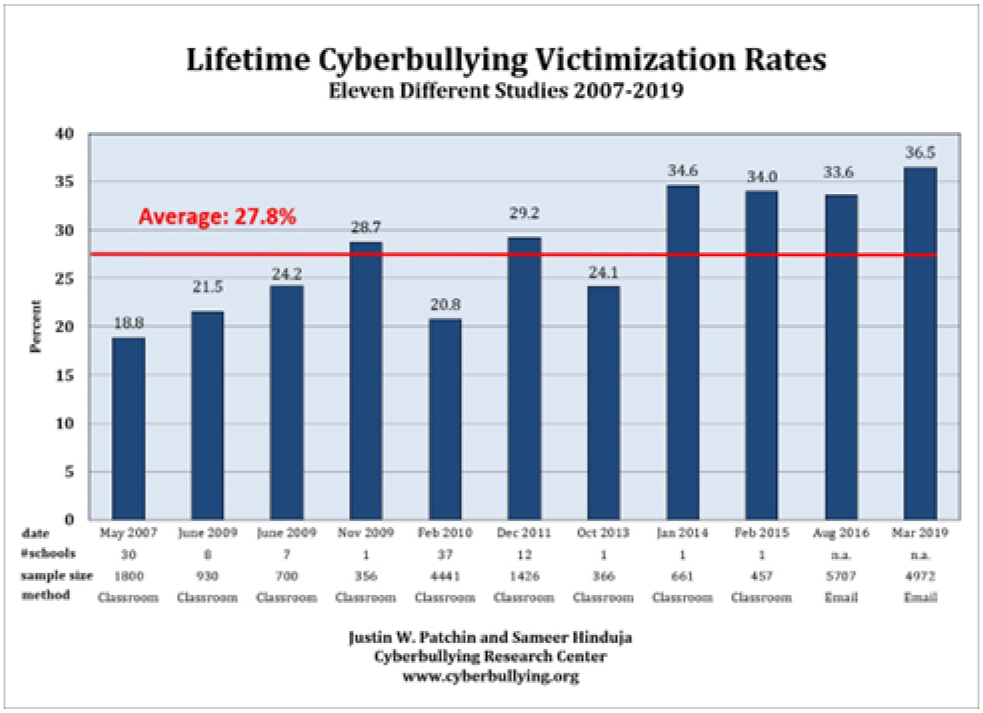
The growth of fraud and scams during the COVID-19 pandemic alone has yielded year-over-year growth inconsistent with past baselines. In fact, the new trend, if continued, would see the scamming industry become a top job opportunity. In 2019, scams in the social media arena reached a damage total of $134 million. In the following year, the same exact figure was reached in six months[23]. Historically, a dramatic shift in the slope of the trend line results in consistent growth with the newfound trajectory, and the graph is beginning to resemble an exponential function. The two platforms, Instagram and Facebook provided an avenue for 7,502 of the 9,832 shopping scams tied to social media, while 1,858 filings did not specify a social media platform. Out of the 43,391 online shopping scams reported to the FTC, 23% were initiated on social media platforms[23]. Instagram and Facebook can roughly equate to a 20% market share of all online shopping scams. It is noteworthy to also point out the pattern of people underreporting fraud cases because of embarrassment or self-blame[25], which leads one to believe that the market is much larger than statistically proven. Thus, the third criteria is properly fulfilled and Scammers are in the S.C.A.R.E. collective.

Chaos Causers

Individuals exist on social media that act “in a deceptive, destructive, or disruptive manner in a social setting on the Internet with no apparent instrumental purpose”[8]. They form a community known as trolls who seek to pollute civilized discourse with incendiary comments and posts. According to a study conducted in 2014, the personality trait of sadism, those who “derive pleasure from cruelty”[8], demonstrated the most robust correlation with online trolling behavior. Investigative journalist Rossalyn Warren corroborates this connection through claims in her book *Targeted and Trolled* that “sadists tend to troll because they enjoy it.”[63] Trolls intend to cause others to suffer so that they can be happy. Directly tying sadism to the act of trolling results is grounds to establish diabolic intent and a win-lose mentality.

The intended outcome is achieved by the troll when their actions successfully disrupt the conventional system of civil communication. The actions taken by a troll do not add value to a society, but rather detract by adding chaos in the form of cruelty. Trolling’s meteoric rise to popularity exhibits the underlying societal transformation in free speech from being a collective entity to an individual one[27]. Transferring the freedom of speech in this manner authorizes a person to ignore responsibility for spewing hate speech under the guise of exercising a First Amendment right. This moral loophole enables an environment of progressively more and more uncompromising dialogue. The greater the adversarial posts and comments, the more people are influenced. Even Google, the owner of YouTube, cited the deleterious effects these practices can lead to by citing internal issues of trolling causing a “chilling effect” on productivity and culture[97]. This cruelty is not benign, as targeted users exhibit increased symptoms of depression, body dysmorphia, and mental health problems. Their thought process and outlook on life were hijacked due to a sadist with a keyboard. A satisfactory manipulation of users is needed for the second criteria.

In addition, statistically significant correlations were discovered between “online commenting frequency, trolling enjoyment, and trolling behavior,”[8] highlighting the effects of excessive social media use. This suggests an inherent structure of social media where one of two possibilities arise (1) high levels of interaction with social media exacerbates an individual’s inclination to become a troll and sadist, and (2) social media is structured in an appealing manner for existing trolls to thrive. Both explanations produce a level of fault on social media platforms. Trolls, or chaos causers, exist independent of platforms, but the role social media plays in enabling these actors to have staggering levels of influence is provable. Figure 5 below exhibits how; cyberbullying, a broader classification of trolling, has grown steadily from 2007 to 2019, alongside social media[72], thus supporting the classification of chaos causers as malicious actors under the third criteria of growth.



date	May 2007	June 2009	June 2009	Nov 2009	Feb 2010	Dec 2011	Oct 2013	Jan 2014	Feb 2015	Aug 2016	Mar 2019
#schools	30	8	7	1	37	12	1	1	1	n.a.	n.a.
sample size	1800	930	700	356	4441	1426	366	661	457	5707	4972
method	Classroom	Classroom	Classroom	Classroom	Classroom	Classroom	Classroom	Classroom	Classroom	Email	Email

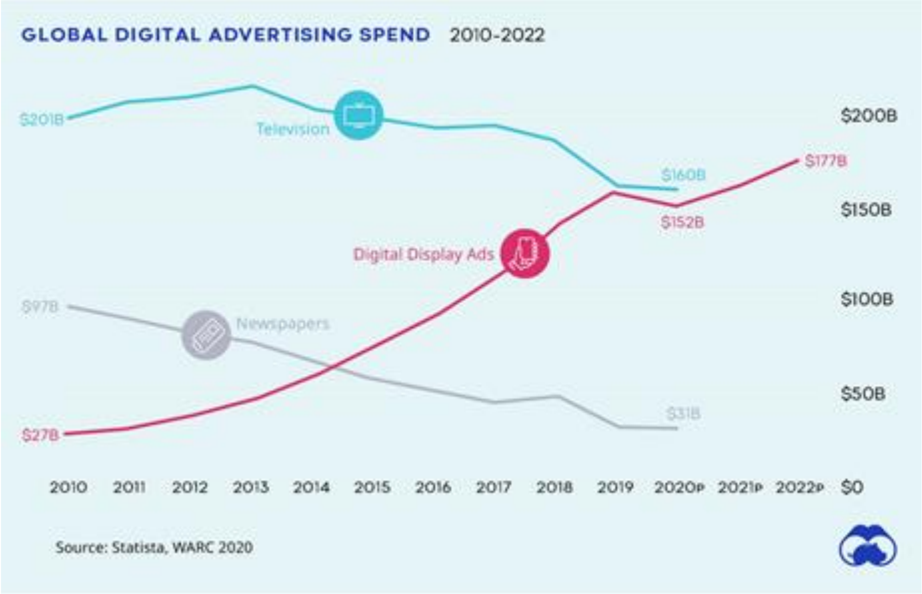
Figure 3. Trolling growth[72].

Advertisers

The focus of social media platforms on advertiser needs and desires over the users has led to questionable collusion between these entities. This advertising-based business model relies on a company paying for advertising space to generate revenue. Higher conversion click rates equate to steeper fees charged for advertisement space[29]. It is well-established that the most effective method is targeted advertisements, which selects ads for users in accordance with data gained from a user’s interaction with a platform, in the form of likes, comments, scroll backs, brief hesitations over an image, search history, and every other recorded piece of data[29]. This generic and seemingly malign categorization was selected for the part played in impacting social media design to generate revenue instead of connection.

Advertisers' negative exploitation and shaping effect on social media deem it the worst of all other actors. The control companies have to persuade Americans is frightening. In 2019, the Statista Research Department polled a sample about the relation their purchases had to them seeing a social media advertisement. An overwhelming 80% of participants reported being compelled to buy an item as a result of an advertisement at least once[28]. The level of influence exerted on consumers by advertisers warrants a malicious classification associated with these actors and a successful manipulation of users.

A colloquial expression goes, "if you are not paying for it[product], then you are the product"[30]. Support for this claim comes from a 2018 estimation of the advertisement revenue Facebook made per each user's data. The average was calculated as $110 per user[17]. However, prioritization of this revenue stream is expected to catapult the worth of the average American's data to exceed $400 by the end of 2022[17]. This growth is directly tied to the effectiveness of targeted advertisements and the development of captology techniques. Social media advertisement dominance is observable by the increase in spending on digital mediums attributed to persuasive techniques producing a higher yield per dollar spent[74]. A suitable level of growth to match that of social media.



GLOBAL DIGITAL ADVERTISING SPEND 2010-2022

Source: Statista, WARC 2020

Figure 4. Digital advertisement growth[74].

Radicalizers

Radicalization is "the process of developing extremist ideologies and beliefs"[31]. Moreover, this establishes a radicalizer as the person encouraging and facilitating this process for another individual. Extremists seek to radicalize for various causes; however, four main groups dominate the literature in the United States: Islamist, far-left, far-right, and single-issue extremists[2]. The process of radicalizing others to act on the behalf of the groups instead of themselves shields their actions from the law. Random victims convinced of the ideology face legal consequences for the acts of terror they were convinced to commit. Therefore, radicalizers have a win-lose mentality by convincing others to conduct dangerous behavior while they remain risk-free.

The best illustration of radicalizers manipulating users comes from a case of radicalization. This fictional account based on the Pizzagate conspiracy theory[1] illustrates a simplified story of a victim of radicalization. The conspiracy theory began during the 2016 presidential election.

Supporters claimed that the United States Democratic officials orchestrated child sex and human trafficking rings in the basements of numerous restaurants. The substantiation originated from allegedly coded messages in emails revealed on Wikileaks. Immediately following the first claims, every credible source refuted its legitimacy, but not everyone listened[SOURCE]. In this case, one minute a person is liking a post on Facebook by their US Representative, and the next minute they find themselves commenting a 1000-word dissertation of outrage on an article detailing the child trafficking ring in a Comet Ping Pong pizzeria basement. The individual stops in the middle of a compelling closing argument. As though coming out of a trance, the individual shakes their head and gawks at the screen in utter disbelief. A racing mind begins frantically searching for the sequence of decisions that led them here, but no clear-cut explanation is found. The above scenario emerges as a frequent occurrence for people worldwide. However, many people do not experience the moment of clarity detailed above. Other accounts of radicalization follow this path, where individuals are unsure of how their thought process transformed from logical to agreeing with extremist theories. This fulfills the second criteria of the manipulation of users.

The database of Profiles of Individual Radicalization in the United States (PIRUS) depicts over 1,800 Americans who were radicalized and subsequently committed acts of extremism from 1948 to 2016[2]. It provides the most comprehensive statistics linking social media use and radical activities. Spanning from religious extremists to any other motivation for committing acts of domestic terrorism, PIRUS represents the first step “from an empirical and scientifically rigorous perspective.” Therefore, it proves the correlating growth of social media and radicalizers. Analyzing the role of social media in radicalization for sequential five-year periods reveals the rapid transference of radicalization efforts to the digital world. In the time between 2005-2010, social media played both a primary role, 1.63%, and a secondary role, 25%, in total cases. The following five years, (2011 to 2016) showed that social media played a primary role in 16.95% of cases(a 900% expansion), and a secondary role in 56.27% of cases(greater than 100% increase)[2]. In 2015, the Subcommittee on National Security noticed these shifts, expressed by the plea that “we must do more to counter the social media threat posed by the Islamic State and other terrorist groups.”[37] Radicalization found an environment suitable for its growth and the acknowledgment of social media’s facilitating role grows widespread. The role of social media in the radicalization trend line’s trajectory can be predicted through isolating datasets found during 2016. There was a social media component in 90% of radicalization instances, an additional 15% from the average of 2011-2016[2]. Future data is expected to corroborate the emergence of social media as a dominant player in radicalization until an equilibrium is reached. The deemed favorite platforms of extremists divulge those they are effective at operating on, and in turn, highlight the platforms to investigate that are conducive to radicalization. To highlight this development, a rapid ascension ensued of social media becoming the dominant medium for radicalization efforts. For comparison purposes, during the 8-year period, 2006-2014, the percentage of U.S. extremists utilizing social media went from 2.86% to 78.57%[2]. Outpacing the adoption rate of the aforementioned poll, concerns are raised that social media’s appeal to extremists is stronger than its appeal to all US citizens[3]. Therefore, significant data supports the growth of radicalizers with social media.

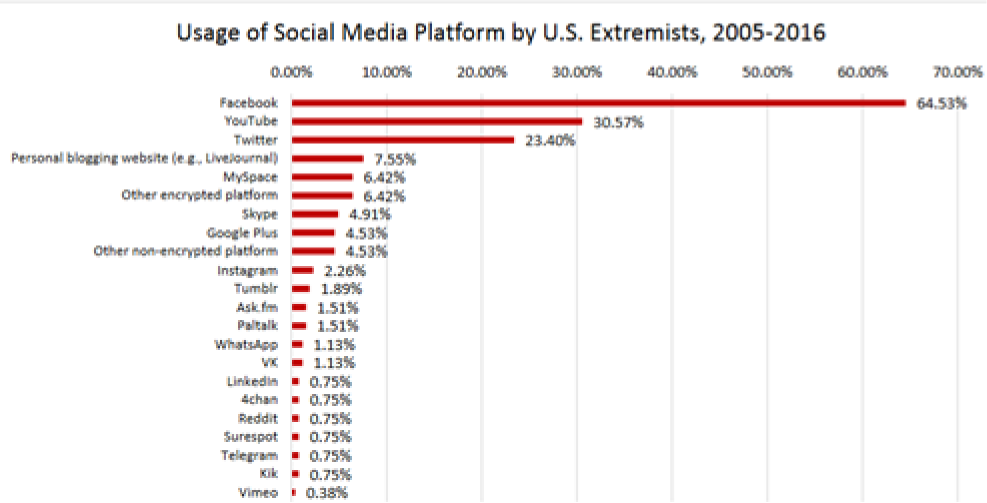


Figure 5. Social media popularity with extremists[2].

Election Interferers

A democratic ideal is not held more sacred than the right of an individual to vote for their representation. Election interference takes aim at this "core institution of democracy"[33] being defined as "one country or group attempting to influence an election in another country"[16]. An expressed intention is to steal the democratic right of individuals and elect an official favorable to the threat actor. If the majority of people voting in an election agreed with the official these interferers wanted, there would be no need to interfere. Therefore, this member fulfills the Win-Loss criteria.

One notable example is Facebook, which emerged as the poster child for the social media industry during the 2016 United States Presidential election and Brexit[10]. Thoughts and beliefs previously deemed extreme had garnered enough support to win votes and elections. This phenomenon is attributed to the widespread use of one entity: social media[10]. It showcased the power the digital world possessed to direct narratives regarding geopolitical issues. In fact, Americans report heightened personal difficulty in recognizing false information on social media since the 2016 election[65]. They expressed the decision-making process relied on to make sense of the world was exploited and damaged because of these campaigns by-election meddlers. Therefore, manipulation of these users occurs. The specific nature may vary, which results in evaluation and judgments on a case-by-case basis. Moreover, the legal ramifications become further clouded when ruling whether an account was producing political content or intending to interfere in an election.

This murkiness stymies a set procedure of identification. Despite this difficulty under international law, it is illegal to interfere in elections based on intention and outcome[33]. However, an expectation of strict adherence to this law should be avoided as historical precedent shows frequent and blatant violations. For instance, the United States and Russia are theorized to have interfered in one out every nine national elections in the latter half of the 20th century[16]. This begs the importance of worrying about interference when an election has a ten percent possibility of being influenced. With the turn of the century came captology and the potential for these two actors to influence every election became more than a fantasy. Once again providing support for Bernays claim of “invisible rulers” in charge of the fate of “millions”[14]. For the requirements of criteria three, Google search trends for Election Interference between 2004-2020 were chosen[73]. This assumes that additional popularity came as the result of increased efforts by election-interfering actors and can be observed in Figure 6. Due to its transitory nature, the catalyzing effect of the online realm on election interference is difficult to demonstrate statistically, therefore, this dataset is sufficient. Prior to the 2010s, minimal people searched for election interference despite it occurring, but in the mid-2010s this amount spiked. It perpetuated a higher level of interest than before.

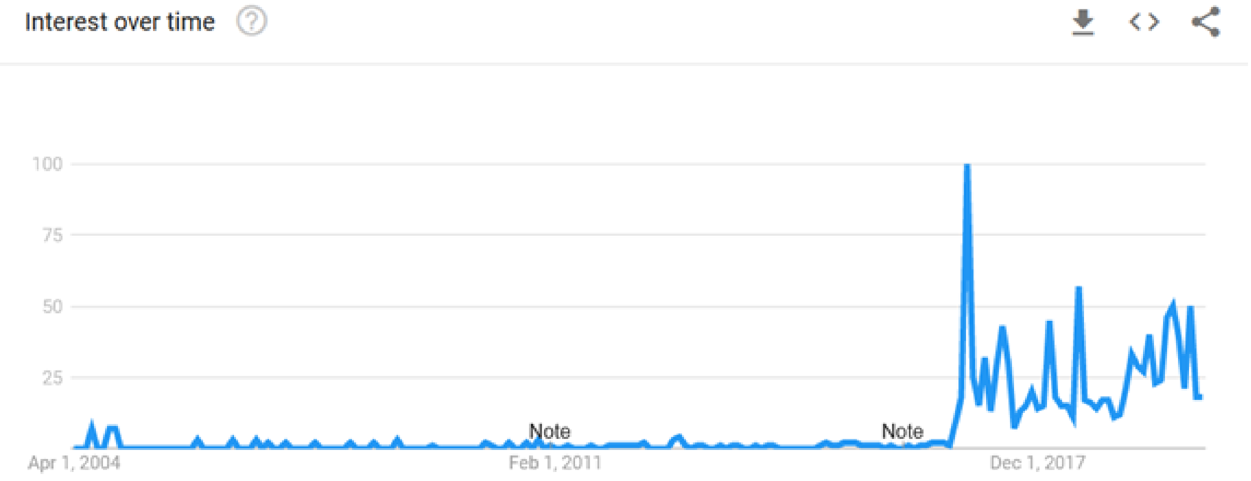


Figure 6. Google search trend of "election interference"[73].

Research Focus

It is the intention of this study to demonstrate their presence is indicative of the platform being hospitable for abuse and fulfills the fourth criteria: exploitation of the platform. Now that the S.C.A.R.E. collective is defined and outlined, it is hypothesized that all actors will exploit social media platforms with a degree of overlap. For this study, the focus stems from this assumption. Therefore, four key questions inform all framing, testing, and analyzing.

1. Do S.C.A.R.E. actors exploit social media platforms for their own gain through similar weaknesses? If so what are these structural flaws?
2. How can these shared weaknesses be evaluated for presence on a platform? Do the created metrics sufficiently evaluate the platforms?
3. Which social media companies display the greatest proclivity of exploitation by S.C.A.R.E. actors? How do these findings compare to prior analyses?

4. What improvements can patch the holes essential for S.C.A.R.E. actors' success? How can this framework be implemented?

Chapter 2

Literature Review

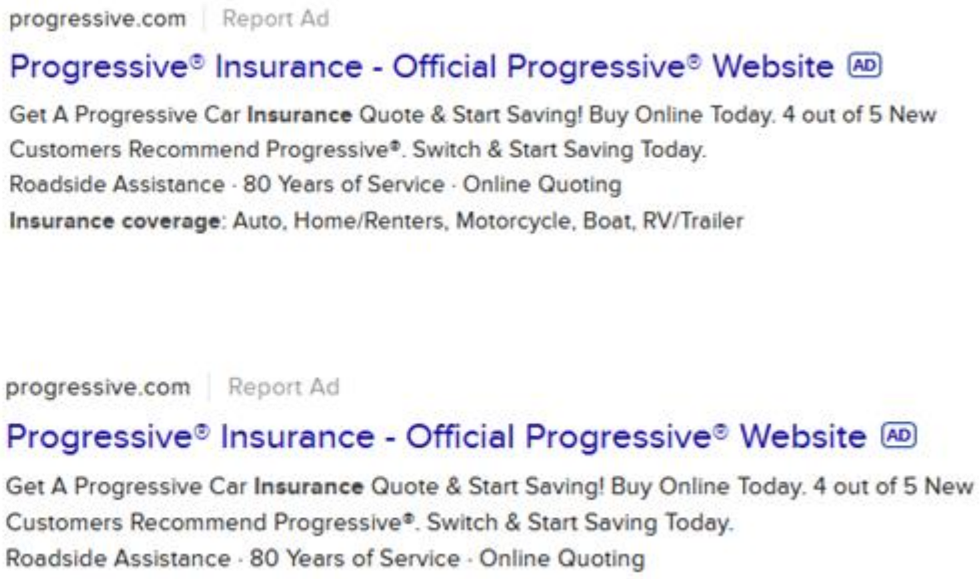
Controlling Mindsets

Since the formation of civilizations there existed an expressed desire to control fellow humans' mindsets. It was a natural occurrence as figures in power relied on people believing in the right narrative and supporting their rule. Consequently, rulers devised increasingly complex methods of control. Catchy phrases and images emerged in more formal ways as society progressed. Documented findings and accounts from the Roman Empire depict the emphasis placed on imperial coin imagery. Emperors inscribed a picture of themselves or a concept instrumental to their rule[12]. For instance, Brutus, a senator that played a role in assassinating Caesar, issued coins with his image on one side and two daggers on the other[103]. After killing Caesar, he found it necessary to gain popular support for future battles. He managed to raise a substantial army partially attributed to this technique, however, his forces were ultimately defeated during the civil war. Messages on coins proved to be an effective method of guiding popular sentiment and explaining policy[12]. Rulers and malicious actors limited by these forms of public relations and propaganda exerted minimal influence. They sought increased power. So nations devoted resources and time to developing more effective techniques. However, no major progress occurred until the 20th century and by a man with the name of Edward Bernays, a nephew of Sigmund Freud.

A massive monetary allocation and technological innovation allowed for future persuasive public advertisements to steer public perception to an unmatched level. Specifically, advancements occurred in persuasive advertisements circa 1910[68]. It was at this time that the purpose of advertisements shifted from informing to creating a realized desire[68]. This evolved objective led to an area of study and research ripe for exploration. A major belief curated by this desire-focused advertisement movement was the notion of acceptable breakfast foods. The idea that breakfast should be the most important meal of the day and specific foods should be eaten are products of cereal advertisement campaigns. James Jackson and John Kellogg were religious men that understood the persuasive guilt that religion possessed. Kellogg's cereal was marketed at sanatoriums on the ideals of health, religion, and hard work, making exaggerated claims, including the cereal's ability to relieve the urge of masturbation[13]. On top of these claims, a sense of maternal duty to serve cereal to children solidified cornflakes' as an American breakfast staple. These marketing techniques solidified breakfast as a moral decision in the minds of Americans and subconsciously "injected attitudes favorable to consumption"[68]. Bernays further exploited the notion of breakfast when contracted to increase bacon sales. His genius strategy consisted of persuading a doctor to issue a statement that a breakfast of bacon and eggs was healthier than a light breakfast. Then, he sent this statement to 5,000 doctors in the form of a petition to gain credibility. Finally, it was published in newspapers and the general public perceived it as a scientific study[13]. The instant success of this newspaper advertisement scheme gained Bernays immense notoriety in the marketing realm. Lucky Strike took notice and hired him to drive their cigarette sales and public interest[69]. After analyzing the disproportionately low sales among women, Bernays crafted an idea. Instead of going directly to the consumer, he convinced fashion and retail companies to empower the narrative that Lucky Strike green was a fashionable color. These companies began displaying window mannequins and magazine models with the color and substantially elevated their inventory of green clothing. Then, he coordinated for a group of women to be photographed smoking cigarettes. These photos appeared everywhere with cigarettes labeled as “torches of freedom”[69]. Capitalizing and partially hijacking the feminist movement, his campaign saw immense success and Lucky Strike cigarette sales skyrocketed. At this point, Bernays received attention from the US government. When the United Fruit company’s interests were threatened in Guatemala, a series of operations started. Now classified under the Banana Wars, the United States launched a coup to overthrow the democratically elected Guatemalan government[69]. Bernays stepped up and launched an information campaign to support the integral nature of United Fruit, explain the interdependence of the US and Guatemala, and criticize the alleged communist rulers, among many fabricated stories to change public sentiment. Once again he was successful and solidified his status as the most influential public puppetmaster. Creating these techniques and putting them into practice gave him insight into the power of influence capabilities[14]. As referenced earlier, Bernays had astute awareness that the world was controlled by those who possessed a mastery of persuasive techniques. After all, Bernays was one of the key players that demonstrated colossal influence in shaping the 20th century[69]. Hundreds of examples are littered throughout history and as other areas of study developed, so did persuasive messaging. Today its’ developments shifted to the online environment and novel methods to exert influence.

Before long, it was discovered that improvements in human psychology and neuroscience could be tested and accelerated in the digital realm with unparalleled ease. The computer allowed for a leap that dwarfed all previous persuasive technology advancements. Focus groups were inefficient and limited by time constraints, while a constant flood of participants waited online. It took until the late 20th century for someone to identify the true power an internet user wielded. B.J. Fogg’s seminal work on persuasive technology, or captology, set the basis for all future research[15]. The outlined benefits which allowed for computers to achieve the desired outcome more effectively included persistence, anonymity, the quantity of data, modalities, scalability, and access[15]. To fully grasp how these shifted capabilities catapulted power dynamics is unfathomable. There was not one ability that improved, but rather more than six. Fogg identified how this immense power forced responsibility on every citizen of the world. Accordingly, captology became a moral debate surrounding the capacity for healthy and positive endeavors to materialize while conversely revealing techniques that facilitate malicious intentions and abuse. Seeking to expand on his previous work, Fogg teaches a course at Stanford University about persuasive technology techniques and systems[19]. Past students are blamed, due to their time in his class, for their instrumental contributions to the design of features, including the like button, red notifications, infinite scrolling, and autoplay videos, along with many others.[19] These features are ubiquitous on highly addictive social media platforms. One can argue that this might be the single most impactful course in the world because of the daily impact these features have on people's lives. Experts expect new methods to surpass these in persuasive capabilities. An incredible concept considering the capabilities platforms already possess. Fresh techniques for persuading online individuals have become a monetarily valuable aim for companies and independent researchers alike.

A popular and effective technique, known as A/B testing, exhibits simplicity in which user preferences can be determined. Implementation of this testing requires a random sample of users, a new feature, and an account for all external factors[20]. Then, half the users are shown the control item while the remaining half view the testing feature on the item. All user data pertaining to interactions with the items are computed and analyzed to upgrade an item's display to maximize the desired outcome, which normally is engagement. This analysis can define causal relationships with strong statistical significance[20]. An example of A/B testing can be seen below. The two progressive advertisements differ in their last line. This variation tests which method is more effective at making a user click.



progressive.com | Report Ad

Progressive® Insurance - Official Progressive® Website AD

Get A Progressive Car **Insurance** Quote & Start Saving! Buy Online Today. 4 out of 5 New Customers Recommend Progressive®. Switch & Start Saving Today.

Roadside Assistance · 80 Years of Service · Online Quoting

Insurance coverage: Auto, Home/Renters, Motorcycle, Boat, RV/Trailer

progressive.com | Report Ad

Progressive® Insurance - Official Progressive® Website AD

Get A Progressive Car **Insurance** Quote & Start Saving! Buy Online Today. 4 out of 5 New Customers Recommend Progressive®. Switch & Start Saving Today.

Roadside Assistance · 80 Years of Service · Online Quoting

Figure 7. A/B testing example[62].

Various other assessments are done on a personal level, building a profile on an individual’s subconscious ideations and motivation. Normalization of these practices informs the present context of the digital world, where a user’s every interaction with the internet is met by a supercomputer on the other end. The more interaction, and user input, an algorithm transforms the subjects/users to act increasingly Pavlovian. Moreover, when given a cue, a person responds in the desired way, ie clicking on an advertisement. The ramifications of opening these windows into the mind have yet to be understood. Leading to ethical dilemmas, Fogg stresses the three obligations facing society regarding captology and the entirety of persuasive technology: amplifying its existence to the general public, combatting those utilizing it for exploitation, and emphasizing the ethical guidelines for innovation[15]. This worry surrounding the importance of society's intervention in persuasive technology was not widespread until recently and consequently few regulations resulted. A grasp of the process through which social media changes user decision-making gives insight into how S.C.A.R.E. actors can exploit users and platforms.

Policies

Social media resides in a limbo status of being a private product or a communal good or service. The debate stems from contention over the level of impact it has on the average American and the ability to narrate discourse in the public sphere. Social media conglomerates went unregulated for a greater part than half of their existence due to uncertainty and a lack of clarity delineating the internet's future path. Consequently, stories of abuse run rampant and accountability is non-existent. If there was not an expectation from the majority of society for regulation this would not be an issue. However, Americans relied on the government to properly regulate these entities and ensure they acted acceptably. Based on cases such as the FDA regulating how cigarette companies can market to children[104]. Sadly the public's trust was misplaced. Social media's insanely fast growth made it too challenging to regulate through bureaucratic means. Speed and evolving entities make policy difficult due to the process and subsequent requirements of regulating something. A regulator must outline the boundaries of its capabilities and current functions. Thus, clearly defining social media becomes vital as the Knight First Amendment Institute at Columbia University states, that to fully grasp how to regulate an entity, the why of regulation must be answered[70]. If this component is neglected, then the resulting regulation will fail or be distorted to serve another purpose. Additionally, the entity targeted by regulations may evade restrictions through a quick pivot of features, practices, and types of abuse. Legislative lag contributed to the environment witnessed in the modern age.

Specifically, the United States lacked the desire in pursuing these entities as they were American-based companies and the economy profited from their creation. Interestingly enough this resulted in the most relevant legal regulation enacted before mainstream social media platforms were invented. Section 230 of the Communications Decency Act of 1996 shields internet companies, namely online service providers, from the liability of actions and speech users engage in on their platform[34].

Section (c)(1) [35]

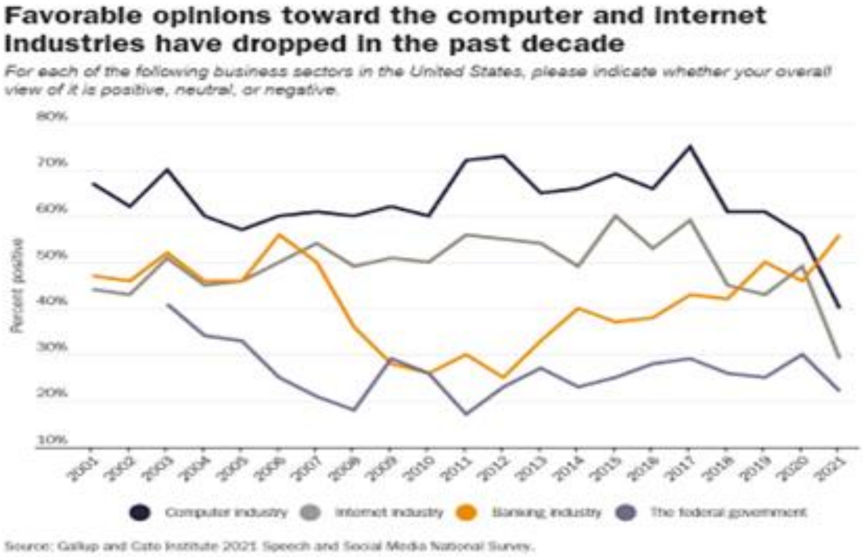
No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.

Section (c)(2) [36]

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

(B) any action taken to enable or make available to information content providers or others the technical means to restrict access to material described in Section(c)(1)

The above clauses serve two primary objectives in providing legal immunity to social media corporations and platforms. Section(c)(1) relinquishes the legal responsibility of content published by another party regardless of any content moderation system the platform has in place. The extent of protection stops if the platform contributed to the content's creation in any capacity[35]. However, enforcement instances of platform contribution do not exist. Then, Section(c)(2) provides the guidelines and legal immunity for which content moderation by companies is permitted on platforms. As well as for the dissemination of content blockers to allow users to self-moderate their personal ecosystem and page[36]. These two clauses are the foundation upon which all US social media content moderation policy is built on. Consequently, the Federal Communications Commission (FCC) stands as the sole enforcer and administer of Section 230. Supporting this jurisdiction, the Supreme Court ruled that the FCC has the jurisdiction to interpret the sections broadly and make modifications when deemed necessary[71]. This designates the FCC as the acting legal authority over determining whether to regulate acceptable standards and practices with a greater degree of complicity placed on social media companies. The FCC's prior experience in regulation covered radio, telephone, and television mediums. Now the government expected it to step up to the major league from the little league. Social media companies won a rigged fight for free reign in the industry. Liability for social media accountability falls on this government agency or more on the lack of government resources to back it up. Beyond a supposed carte blanche held by the FCC, the landscape of policy and regulation for companies is nonexistent.



Favorable opinions toward the computer and internet industries have dropped in the past decade

For each of the following business sectors in the United States, please indicate whether your overall view of it is positive, neutral, or negative.

Source: Gallup and Cato Institute 2021 Speech and Social Media National Survey.

Figure 8. Public opinion toward industries[22].

Since the dawn of social media, the United States Government has been wary about issuing legislation that would regulate these entities due to the topic's partisan nature in encroaching First Amendment rights. Social media conglomerates are put in a similar situation, with any decision potentially alienating a chunk of their userbase. Their lack of standard policy and repeat abuses has resulted in declining approval ratings of the internet and computer industries[22]. While approval ratings of the government have decreased as well, the multitude of factors beyond just social media policy is substantial. Social media management shares partial liability. This outlines the hardest element of the situation society finds itself in. No one party is completely responsible and everyone wants a single scapegoat. While being a consideration of this study, it exceeds the outlined scope. The focus revolves around the key structural flaws that grew because of this colossal failure of government, S.C.A.R.E. abusers, social media companies, and others.

A brief background of other nations' social media policies serves to illustrate the reasons why the United States is the primary subject of analysis. Further, it conveys the global issue this threat becomes the longer society ignores the root causes and tries for ineffective policy. First comparing the policies on content moderation in the United States versus other nations proves revealing. In 2017, the German government passed a bill, dubbed NetzDG, intended to hold social media platforms accountable for failing to properly moderate content. The primary reasons grew from the proliferation of hate speech originating in Neo-Nazi and radical groups. Attempting to take a strong stance, the government included harsh penalties. In the worst cases, fines amount to sums greater than 50 million Euros[85]. An overt attempt was made by the government to shift blame onto social media companies, contrasting heavily with the Section 230 immunity guaranteed by the United States FCC. However, the practice described as "notice and takedown" in the German policy receives year-over-year backlash for ineffective enforcement. Only a few trickle-down reports and cases occurred in the following three years[107]. Even harsh penalties and bipartisan policy find difficulty in making a difference. One begins to wonder if the way policies take aim at abuses is unfounded. Instead of catching the water out of a sprinkler, one turns down the water flow from the hose. In the same way for content moderation, the comments prove impossible to stop, but when the perspective shifts to the primary reasons for content moderation, a solution seems feasible. Again patching social media's structural flaws surpasses policy when evaluating logical and practical effectiveness.

On the other side of the world, Australia has taken steps to fight trolls and defamatory content with its proposed Social Media (Anti-Trolling) Bill[39]. This bill's development was in response to a contentious court ruling that assigned partial liability to users whose pages were the subject of anonymous defamation posts and hateful content. Regardless of the user's level of awareness, legal officials may issue fines[39]. The Australian government noticed this violation and decided action was necessary. The aforementioned proposed bill legally compelled social media platforms to supply the contact information of an anonymous poster in a defamation complaint under an unmasking clause. If the platform does not or cannot provide the required contact details, the service provider will be treated as the publisher of the defamatory post[39]. Then, parties or regulating bodies file complaints against the platforms for negligence, and fines result. A unique approach that takes aim at anonymity and tests the limits to which social media companies protect their users. The hypothesized benefits of the bill passing are reduced reported cases of defamatory content, improved moderation by service providers, and a more congenial environment to engage in social networking. Critics state this bill will fail due to the lack of “evidence-based” frameworks to identify this behavior. Instead, they suppose the results will be wasted money and court time of judges[108]. As seen in the previous content moderation policy with Germany, a policy to punish offenders neglects to consistently and firmly outline what actions deem an individual as an offender. Referring back to the Knight Amendment’s determination of how to regulate an entity, both these policies are utter failures[70]. Admittedly, this shows a step in the right direction. Precedents, as displayed here, represent a shift in the attitudes governments and legislative bodies establish toward social networking. For example, the stark contrast between this bill and Germany’s Network Enforcement Act with Section 230 Immunity Clause leads to a rift in policy decisions between nations. Eventually, this could force social media platforms to accommodate differing policies and from this chaos better methods every country agrees on will emerge. Both Australia's and Germany’s legislatures seek to raise and diversify the minimum standards service providers must follow to operate as a global platform.

The gap between what regulations should be and are is massive. If the government tried to close this chasm when it was a crack, regulation could have succeeded. In the present day, policy shows minimal impact and society requires an outline of improvements for future platforms. This study provides the basis for a framework as regulating bodies utterly failed to meet this requirement.

For this research, the United States serves as the primary subject due to four reasons: the lack of policy or action, the origin of most social media platforms, and internet penetration rates. Comparably, the United States does not display the similar desire of Australia and Germany to protect their citizens from malicious actors. Additionally, the most famous cases of social media abuse originate from the United States. While there are some notable exceptions, these do not compare to the sheer magnitude of cases in America. Next, the four most popular social media platforms: Facebook, Instagram, WhatsApp, and Youtube; all originated in the United States[109]. Finally, the United States has some of the highest internet and social media penetration rates of all nations[114]. The combination of the above four reasons solidifies the selection of the United States as the primary country for evaluation in this study.

Previous Frameworks

Before this study began, the focus was different. The original intention was to use existing frameworks rating social media abuse in order to create a secure online platform where users could operate freely and safely on. Preliminary research proved this aim was infeasible. There was no definitive existing literature on or closely related to what the author sought. Some studies identified flaws in social networking and others detailed how S.C.A.R.E. actors achieve success, albeit not the identical actors grouped together in the same way. However, these frameworks supposed the flaws as unchangeable, something that society has to learn to live with if it wants the benefits. Due to no prior frameworks for categorization of a collective, like S.C.A.R.E., that abuses social media, nor for rating social media's likelihood of being exploited and abused, the primary purpose of this study transformed. While prior research is limited there exist frameworks containing information and helpful ideas. The below frameworks frame the online ecosystem and extremism in a unique manner. Select issues and factors are highlighted, yet they stop short of outlining any comprehensive system. They provide great insight and topics to consider in order to create a methodology for grading and categorizing S.C.A.R.E.'s exploitation of social media.

Violent Extremism Evaluation Measurement

The Violent Extremism Evaluation Measurement is a general framework of warning signs detailing how a user is abused. It gives an overarching summary of the process from beginning to end. The framework depicts the forming characteristics of extremism, the first signs, and final outcomes. These stages are depicted as the “Initial states of extremism”, “Initial manifestations of extremism”, and “Extremist manifestations”. Within each stage, specific inclinations and actions highlighted the progress of an individual on the extremist path. The pertinent phase for this study is “Initial manifestations of extremism”, which describes the qualities that are symbolic of a person becoming radicalized. These consist of the following: “Posting, sharing, and interacting with extremist content in social media”, “Regularly accessing and viewing extremist websites and engaging with others on extremist forums and chatrooms”, and “Identification with, belief in, and acceptance of extremist narratives”[105]. From these three traits, one extrapolates the components of social media that support extremism and radicalization. Understanding these three initial manifestations, one interprets the beginning of radicalization as a person interacting with the content, and community, then shifting their beliefs to fall in line with radical concepts. So, it becomes crucial to identify the qualities which increase the likelihood of each of these factors. Specifically, what medium or platform provides the greatest alignment. As shown in the reasoning for radicalizers as S.C.A.R.E. actors, social media is that medium. It allows for the greatest growth and the migration of radicalizers occurred as a result. The importance of this framework lies in its definition of how to gauge a S.C.A.R.E. actor's initial success and impact. However, the case studies chapter explains the specific methods by which radicalizers achieve success including exploited weaknesses. The VEEM framework informs the deeper study of this one S.C.A.R.E. actor and favorable components sought. Based on the limited relevance of preliminary frameworks of each S.C.A.R.E. actor, the author chose to detail VEEM in order to illustrate the way they exceed the scope of this study.

The Honeycomb Framework

Another framework, named the Honeycomb Framework, describes the basis of how someone socializes online. It seeks to display the tradeoff between positive social media functionality and the resulting negative counterparts. It expresses how the initial public conception and media presentation of the benefits were positive, but now it dwindles every year. The authors contend this was predictable. A combination of functions social media achieves directly corresponds to abuse. These functions are covered by seven groups in which an individual exists online. Social media's functions for an individual consist of sharing, presence, conversations, identity, relationships, groups, and reputation. All abuse on social media derives from these seven functions. Abuses outlined are election meddling, disseminating fake news, trolling, intellectual property leaking, and numerous more. A basic overview of each function and its dark side covers the necessary knowledge to comprehend the implications of this framework in the ensuing analysis. The first functionality of sharing provides massive amounts of information to every user in the form of user-generated content(UGC), but inappropriate UGC spreads just as rapidly and widely. Then, a user’s presence indicates how accessible a user is and where they are, however, this leads to nefarious location tracking and monitoring. Next, conversations describe how users communicate with one another, yet not all this dialogue is positive with misinformation, disinformation, and aggressive engagement. The center of the honeycomb structure is an individual’s identity. Identity is a person’s online representation and the pivotal component of a social media profile, therefore, the target of exploitation by outside actors. This exploitation occurs through the fifth functionality of relationships: the extent to which users can connect to each other. Primary methods of abuse are threat, coercion, abuse, and intimidation of the user. Through continued abuse, a person’s reputation, the sixth honeycomb component, takes a hit with shaming and defamation. Finally, the interaction of an individual with a group leads to echo chambers and biases when processing information[106]. Overall, this framework aptly analyzes the expression of a real person in an online environment.

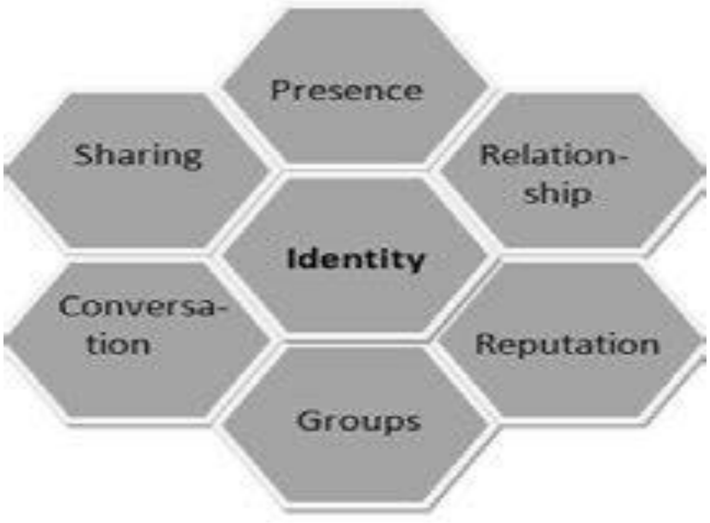


Figure 9. The Honeycomb Framework[106].

Copying the online world to a digital landscape does not come problem-free and the greatest functions, double as flaws. In certain aspects, these issues destroy the foundations of human socialization and drive the species further apart. The Honeycomb framework highlights these potential abuses to an apt degree. Its main applicable tenets involve the dark side where functions, or features, of social media, connect to malicious actions and outcomes. Relevance to the current analyses requires an approach of reverse engineering. This means starting at the abuses and following them to the root cause, the functions, then connecting the functions to features social media platforms contain. Improperly implemented functions result in flawed features, which set the stage for abuse. Features are the intermediary as the functions in and of themselves are not evil, but dealt with incorrectly and malicious actors pounce. The missing element of the honeycomb framework demonstrates the need for in-depth exploration of how abusers operate on social media platforms. Specifically, the features that allow for their success. Not all these functions and abuses fall within the scope of the thesis. Therefore, the S.C.A.R.E. actors and their actions decide which functions to include based on the abuses led to as depicted by the Honeycomb framework. Organizing these functions and abuses in a table format, allows for additional connections. These include identifying the S.C.A.R.E. actors that engage in each type of abuse. Further, the case studies section will provide the features that lead to the function being abused in a specific manner.

Function	Feature	Abuse Type	S.C.A.R.E. Actor
Sharing	TBD	Inappropriate UGC	S.C.A.R.E.
Groups	TBD	Echo-chambers/Bias	S.C.A.R.E.
Conversation	TBD	Misinformation/Disinformation/Aggressive engagement	S.C.A.R.E.
Identity	TBD	Target of exploitation (user)	S.C.-.R.E.
Relationships	TBD	Threat/Coercion/Abuse/Intimidation	S.C.A.R.E.
Reputation	NA	Shaming/Defamation	-.C.-.-.-.
Presence	NA	Location tracking/Monitoring	-.-.-.-.-.

Table 1. The Honeycomb framework and S.C.A.R.E. actor analysis

The seven building blocks inform this socialization process, which hopefully companies utilize to improve platform design going forward. A constant reminder of this honeycomb framework helps to establish a structural analysis of how these functions translate to features S.C.A.R.E. actors exploit and in turn lead to types of abuse. This represents one influence on the case studies component of this thesis.

Chapter 3

Case Studies

It is the intention of this study to demonstrate that the presence of structural weaknesses on a platform indicates the platform is hospitable for abuse and fulfills the fourth criteria: exploitation of the platform.

Scamming

As stated previously, scamming on social media platforms has achieved an exponential rate of growth. The fundamental components enabling this pattern are exposed when analyzing popular cases of social commerce fraud. A routine case of social media scamming starts rather simply. A seemingly real user, who is the scammer, contacts an individual. The scammer advertises a product or starts a dialogue with the individual. Then, the scammer compels the victim to buy a product, send money, or give information. Finally, this scammer stops communicating with the user and fails to deliver on their end of the agreement. If this final step occurs differently, no scam took place. This component makes detecting scams difficult as the adversary has a considerable amount of time between the user sending them money, or information, and them realizing a scam occurred. The relatively uneventful nature of this S.C.A.R.E. actor differentiates it from the traceable signs the rest of the actors give.

A person’s whole identity exists online and this includes names, date of birth, pet names, and phone numbers. It allows for an atmosphere of new connections and real-life ones continued in this digital landscape. However, it also opens the door to scammers who prey on this level of personalization. Essentially, the more information known about the user the better content recommendations are, which results in the user trusting the information as it builds in their own biases. Analyzing the dark side proves this personalization serves another purpose. In the case of identity theft scams, scammers need only three pieces of the listed information about a person's identity[112]. This becomes even easier for scammers when platforms improperly handle their users' information. Then, emails, phone numbers, etc, become public information ripe for the picking and abuse from waiting scammers. Examples include Facebook leaking the information of over 6 million users in 2012 and Twitter revealing the information of 1 million of their users in 2013[112]. These are two cases from a vast pattern of repeated inadequacy by those companies charged with protecting online identities. Both data handling and personalization open users to vulnerabilities from scammers. It appears that driving engagement up also drives scamming success.

According to a study conducted on Malaysian participants, there are qualities that determine online fraud's success, particularly on a social networking platform[76]. Taking into account the responses of 707 social media users, they analyzed factors including the purchase of goods, social media use, and prior scamming incidents, among other data points. The authors concluded that there was a positive relationship between the length of time spent on a social networking platform and the inclination to make a purchase[76]. Platforms seek this correlation with their advertisement engagement objective. A user's duration spent on a platform is influenced heavily by engagement features utilizing captology to hold and keep a user's attention. Furthermore, the likelihood of being scammed escalates with the number of purchases made online. This suggests that increased social media use leads to a greater probability of becoming a scam victim[76]. Although engagement in and of itself does lead to scams, the study gives additional credit to social commerce's features of "anonymity, non-transparency, invisibility, and payment method”[76]. They declared with certainty these features played a role in enabling scammers. In this sense, the platforms appear to corral their users akin to fish for scammers to throw nets at. Albeit this is not the situation or the initial intention, it manifested as such. The objective matters less when the impact reaches this scale of abuse. If systems created guidelines to identify these nefarious actors, the platforms would gain partial praise for fighting the issue. Unfortunately, the sheer volume of scams creates an insurmountable challenge.

Moderating fake advertisements, accounts, or previously reported scams, becomes paramount. The procedure for identifying the characteristics of a scammer and subsequent moderation became disjointed over time. The context of content moderation and current techniques provides a comprehensive picture of the problems today. While research traditionally details the content moderators located in the United States, the majority of moderating has been exported to India among other nations in “the Global South”[77]. The Global South refers to less-developed or developing nations that have lower economic and geopolitical power in the global ecosystem. Barriers emerge in the sheer volume of content and comprehending complexities of accepted behavior, therefore, moderators share the workload with automated processes. However, automated systems express immense difficulty in recognizing slight deviations in what appear to be legitimate profiles at first look, according to the founder of a top Indian content moderation firm[77]. The combination of Indian moderators' analytical ability to assess scams and automated algorithms' capability of working through vast amounts of content appears to be a solid method. It seeks to address the totality of incongruencies each moderating entity faces. Unfortunately, the separation of analytical human skills and quantitative computing leaves a significant portion of scams free from moderation. Another difficulty emerges in anonymity preventing scammers from capture. Once one account receives a ban or is shut down, another account immediately surfaces.

In fact, fake accounts exceed 15%, 48 million, of all Twitter users[112]. By the time an individual realizes they fell victim to a scam and reports it, the scammer already deleted or moved on from the account used. Constantly shifting aliases makes social media a S.C.A.R.E. actors' dream world. The demonstrated shortcomings shed light on the core features enabling Scammers. Definitively outlining that attribution of success to eight structural features: engagement, invisibility, data handling, non-transparency, payment method, content moderation, personalization, and anonymity.

Structural Weaknesses: 1) engagement structure 2) personalization 3) anonymity 4) invisibility 5) data handling 6) non-transparency 7) payment method 8) content moderation

Chaos Causing/Trolling

The digital manifestation of sadism, better known as Trolling, has caused self-loathing, anger, and many negative emotions. These incendiary results are based on and achieved through their vitriolic nature online. In practice, trolls camouflage themselves within the larger digital ecosystem. After all, how can actions consisting of a user posting an “inflammatory” comment on another’s post and “sharing inappropriate content” on their page be differentiated from the acceptable behavior of users[78]? Despite this lack of certainty in defining what speech or actions indict a person as a troll, the impact is visible. Referred to by countless synonyms-such as a bully, antagonizer, provocateur, cyberbully, provoker, and last but not least, troll, this categorization of S.C.A.R.E. actor thrives on the framework of structural weaknesses on social networking sites. A narrative example and commentary by a Twitter executive articulate these damaging real-world effects articulated best.

This tragic tale recounts the story of a 17-year-old boy named Felix Alexander. A case that gives observers insight into how trolls operate and the truly devastating impact they have on communities[81]. It is a prime example as the facts are straightforward without complexity obstructing an analyzer's understanding. At the age of 14, Felix was an avid social media user. Much like the rest of his generation's migration to social networking, its importance started to rival real-life interactions. A couple of months into his 14th year, he began receiving nasty messages. As days turned into weeks, these comments grew to a level impossible to ignore or solve with a simple block. To give context, the following insults were included in these comments: "black rat", "ugly", "worthless", and "everybody hates you"[81]. The majority of these keyboard-warriors comprised a group of people that never met or remotely knew Felix[80]. Rossalyn Warren, in *Targeted and Trolled*, states this trend occurs frequently in cases as "trolls take advantage of the freedom to be cruel to strangers without their identity becoming known, so there's no accountability, and without having to look their victim in the eye."[63] This reveals a flaw in the identity function not highlighted in the dark side of the Honeycomb framework. The flaw lies in the lack of systems for non-repudiation and authentication. Unfortunately, this hate only became larger with time. It overflowed to his classmates taking part and bullying/trolling him. It hopes of starting fresh Felix and his parents decided to move schools. However, social media did not care. These trolls continued to make his life miserable even after switching schools. With comments daily, Felix's feed turned into an echo chamber of disparaging and insulting content curated just for him. This group that forced Felix to join, gave him only negative content, but at the very least he was connected. For a duration of time, he made a brave decision and completely deleted all social media applications. This was short-lived. His mom commented on feelings of isolation and deleterious withdrawal symptoms. Felix's apparent need for social media to engage with and be a part of a group even if those in the group were horrendous human beings. A parallel to hard drug users is undeniable at this point in the case study. This association reveals the addictive qualities: Felix would rather suffer at the fingers of randoms than be alone. Tragically, Felix saw no options to escape from this torment. In 2016, he took one small step onto a train track and ended the misery forever[80]. The purpose of this story goes beyond a breakdown of the structural flaws of anonymity or engagement and speaks to the dire consequences of ignoring these S.C.A.R.E. actors that destroy and end lives. The case above represents the extreme of how damaging trolling behavior and chaos causers are.

This paints a grim picture for the next generation. In fact, 70% of adolescents between 13-22 have been cyberbullied or trolled[84]. Not all of these cases result in suicide, but that does not diminish the problem of social media issues being endemic. Due to this severity, the issue of moderating posts comes under scrutiny. In an open letter, Felix's mother highlights a core issue present when platforms decide whether to moderate this type of content. She claims that while a passive entity does not actively participate in bullying, they serve in an enabling capacity "by not reporting it" or supporting "the child being abused, which just validates the bully's behavior"[80]. If accounting for social media platforms as passive entities, the implications would find platforms liable under negligence to act. Understandably the platform has legal impunity, citing Section 230(c)(1)[35], and the likelihood of the FCC taking a stand against the companies is minimal. With a basic contextual understanding of the Felix story, one may simply infer that society holds social networking companies to a lower standard than teenage boys[80]. Platforms may claim it is out of their power to moderate harmful content, but a former head of a major social platform tends to disagree. As articulated by retired Twitter CEO Dick Costolo, before leaving twitter, in his famous statement regarding content moderation on the platform: "We suck at dealing with abuse and trolls on the platform and we've sucked at it for years…We lose core user after core user by not addressing simple trolling issues that they face every day."[83] Immediately following this statement he admitted "full responsibility" and claimed all resources necessary would be tasked to handle the issue with "responsibility and accountability"[83]. This declaration was made in 2015, a full year before Felix committed suicide, in the midst of the Russian Twitter troll campaign discussed in a later case study[61], and four months before quitting[82]. A different narrative of helplessness presents a new lens to draw up opinions around content moderation, or the lack thereof. As evident in the discussed cases, vital issues have resulted in widespread and acknowledged abuse of platforms by trolls. These structural weaknesses consist of anonymity, engagement, echo chambers, and content moderation.

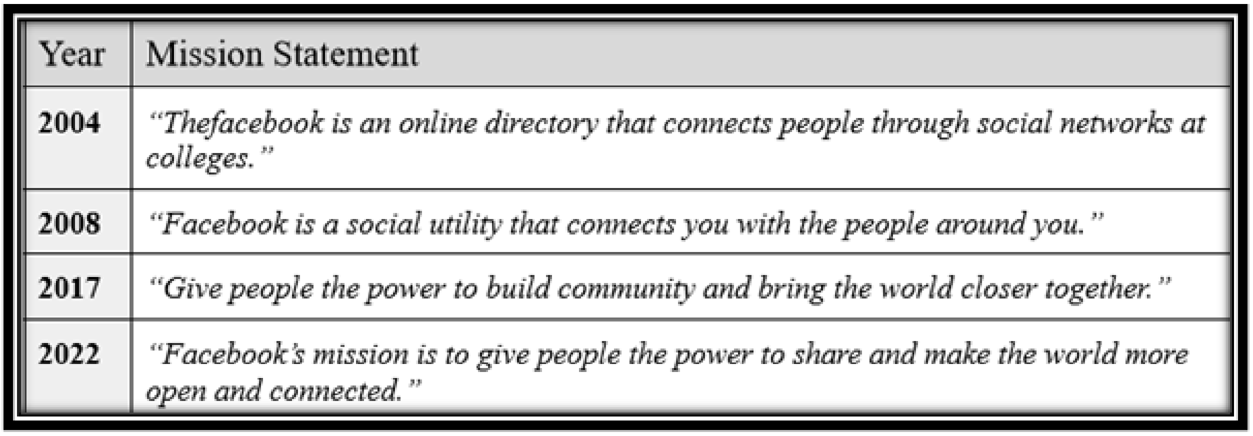
Structural Weaknesses: 1) anonymity 2) content moderation 3) engagement structure 4) personalization

Advertising

The role played by advertisers in abusing social media's features for profit is seen in a unique way as opposed to the other four S.C.A.R.E. actors. The contrast lies in the monetary relationship with social media companies. In fact, the case study of Facebook demonstrates how the platform shifted key components when the platform was still largely mutable to better accommodate this S.C.A.R.E. entity. The precedents set by Facebook can be traced as present in several social media platforms and still altering newer platforms.

Facebook received a tremendous wave of scrutiny for the risks it invited based on its main features, specifically relating to the core tenets boosting its success. A paper from 2009 broke down Facebook and social media as a whole’s attraction to users[5]. It states Facebook’s appeal is based on three theories: the “uses and gratifications” theory, the “third-person effect” approach, and the idea of “ritualized media use.”[5] Essentially, the first theory depicts individuals' inherent need for identity construction, entertainment and distractions, and relationships. The second theory communicates that users tend to associate the adverse effects of something with others while transcribing the positive ones to themselves. Finally, the third depicts the subconscious entrance of social media into people's daily routines[5]. When simplified, the structure of Facebook is built around a need individuals have, assuring their downplaying of risks, and forming a habit of user engagement with the platform. The report goes on to highlight serious flaws in Facebook and commentary by the “watchdog organization Privacy International that charged Facebook with severe privacy flaws”[5]. The major concerns observed by Privacy International surrounded how user data was handled insecurely and distributed openly to anyone who wanted it. This analysis took place six years after Facebook’s conception and set the stage for the future. This habit of user engagement presents a rare opportunity for a certain group of actors.

Acknowledging this base theory of operation, Facebook’s potential as a persuasive technology became evident to advertising companies. Any product or idea an agency wanted to be sold could find a highly engaged audience assured to interact with it. The evolution of Facebook’s mission statement displays this shift in purpose from user to advertiser centric through subtle wording modifications[18].



Year	Mission Statement
2004	*"Thefacebook is an online directory that connects people through social networks at colleges."*
2008	*"Facebook is a social utility that connects you with the people around you."*
2017	*"Give people the power to build community and bring the world closer together."*
2022	*"Facebook's mission is to give people the power to share and make the world more open and connected."*

Table 2. Facebook mission statement progression[18].

Initially, Facebook has a pure mission and demonstrates the massive possibility users now possessed. Facebook went public in 2012, and the implications are drawn when contrasting the mission statements of 2008 and 2017. No documented Facebook mission statement updates occurred from 2008 until 2017. Although sources debate there were numerous revisions, moreover, they disagreed with what the 2017 mission statement actually stated. This confusion escalated with releases by Facebook detailing conflicting mission statements. This change showcased, whether intentionally or not, that the platform's sole concern was not connecting people to people. Additionally, the Cambridge Analytica scandal immediately preceded the change of mission statement in 2017[38]. This forced Facebook to take a broader stance in order to avoid liability. In comparing the 2004 and 2008 missions with 2022, a connection can be read as the goal for each, but who Facebook intends to connect its users to becomes open to interpretation in the latter. The description best corroborating the details mentioned earlier around Facebook's design is the connection of advertisers with users. They were revealing the true dynamic of customer-product relations as an inherent dependence formed: Facebook must prioritize advertisers' needs. Giving people the power to share allows for more information to be recorded about the individuals. The information age sees this user information traded for revenue from advertisers. While the true depth in which Facebook serves advertisers' interests is barely exposed through mission statement analysis, company memos and outside reports confirm it.

The prioritization manifested in unsavory outcomes as the needs of advertising companies are aligned closely with the others four S.C.A.R.E. actors. In a 2019 internal Facebook memo, an executive stated: “We also have compelling evidence that our core product mechanics, such as vitality, recommendations, and optimizing for engagement, are a significant part of why these types of speech flourish on the platform.”[6]. This was the response to why hate speech spreads successfully on the platform. This points to the growth of the structural features: engagement, vitality, and personalization, present on Facebook as a result of fine-tuning its mechanisms for advertiser benefit. Ingrained at a level near impossible to remove without heavy losses befalling the organization. However, content moderation could hypothetically slow the disastrous effects of these features. Until the realization hits that content moderation threatens the bottom line profit. Essentially, Facebook developed from the three primary reasons for success[5] to serve users, then advertisers took interest and Facebook refined its captology techniques. These primary reasons generated the highly impactful “core product mechanics”[6] that reward misinformation, hate speech, and abuse. If content moderation became stricter, the impressive levels of engagement and interaction would fall.

Assessing blame is difficult. The priorities advertisers forced onto a young and complacent Facebook is either textbook grooming or selling out for profit. Advertisers' part played in advancing their own agenda while leaving the door open for a collective of malicious actors matched only by Facebook’s complicity. Their influence can best be categorized as malicious due to their damaging impact on social networking. When generating revenue becomes dependent on monetizing the personal information of a user base, the platform is no longer social networking, it is advertising networking.

Structural Weaknesses: 1) engagement priority 2) personalization 3) vitality 4) content moderation 5) rate of misinformation spread 6) data handling

Radicalizing

Radicalization existed long before social media, however, as demonstrated earlier, it received a catalyst for growth in the form of social networking. The migration of radicalizers to platforms outpaced the average human[3]. So, what specific mechanisms were attractive enough to focus efforts on this digital medium. Moreover, which features encouraged the radicalization of individuals who would have not sought out extremist groups. Now being exposed to this content in such an appealing and persuasive manner that they can not resist. This case study reveals the root causes and enabling features through referencing the discussion on persuasive technology, examining the process by which radicalization occurs, and studying statistics and internal reports from Facebook, Twitter, and Instagram.

Any individual has the potential to be radicalized. The process follows the psychological principles of constant exposure[95]. A person constantly exposed to graphic content leads their brain into a state of emotional desensitization and mortality salience, a state of being overwhelmed by the thought of one’s own death. This combination legitimizes extremist actions in the individual’s head. Eventually, leading to a heightened willingness to commit radical acts of terrorism such as suicide bombings and violent assaults[95]. The constant exposure resembles an echo-chamber environment where a user interacts with extremist content once and then it slowly dominates a user’s feed. This level of personalization makes the user feel a part of a community and as any person does to fit in: the person changes their perception. Personalized social media content makes extremism more appealing as a false reality forms in the user's head. A new reality that paints radicals as “positive and desirable”[95]. Instead of disgust, the user’s brain starts to rationalize these actions and identify with them. In turn, this causes quicker adoption rates by non-extreme citizens through a simplistic process, credit given to social networking platforms.

Online platforms such as Facebook and Twitter hastened online radicalization and the time frame of radicalization in general. Observation of radicalization duration corroborates this claim by a decrease in the mean time of radicalization from 18 months to 13 months over the span of time, 2005-2016[2]. A 28% decrease in the time needed to radicalize one person. Companies claim this is an unfortunate cost associated with the level of connectivity they offer. However, upon looker deeper, indisputable proof of platforms’ complicity emerges. Specifically, in regard to the algorithms tasked with recommending topics and information to users. An internal Facebook research report conducted in 2016 concluded that 64% of all people who join extremist groups were found to have been recommended the group[94]. Whether the person was already radicalized or not, does not matter in the slightest. Facebook endorsed radicalization by recommending a user join an extremist group in the majority of all cases on the platform. Despite this memo making Facebook executives aware, nothing is done. The inherent structural flaw creates this horrible byproduct. It is a feature so intertwined with the monetization of the platform that it is invincible. The feature is named engagement. Trained machine learning models suggest information that will engage a user. At a preliminary level, this makes logical sense. Unfortunately, conducted by MIT, discovered that humans migrate toward exciting and catchy headlines; thus, this type of speech and misinformation spread anywhere from 6 to 20 times faster than facts[7]. A Soviet-era saying expresses this phenomenon best. There were two primary newspapers from 1917-1991 in the USSR, *pravda*(“truth”) and *izvestia*(“news”). One was issued by the communist party and the other by the government. The joke went “*Neito Pravda uv Izvestia a uv Izvestia nieto pravda”* or in English *“*there is no truth in the news and no news in the truth.”[110] For this reason, attention-grabbing misinformation becomes the dominant posting preference of radicalizers. Ultimately, they achieve success through a crafted recommendation by Facebook’s algorithms which engineer the user for this result through constant A/B testing and priming[20] Platforms favor false and misleading content. The next logical step witnesses platforms moderate this content to a stricter degree, but that decreases engagement levels as the truth can not compete with fake news.

If properly moderated this psychological phenomenon and outcome would be limited, but the expressed failure to identify and block radicalizers on platforms becomes apparent. In 2014 ISIS members were estimated to have generated 200,000 pieces of content per day on Twitter[95]. The number of accounts suspended by Twitter showed an attempt to quell this content until a closer look is taken. Specifically, one ISIS member is acknowledged to be responsible for over 100 suspended accounts with many more unidentified profiles expected[95]. Fake account creation appears to negate any moderation efforts taken by platforms as nothing directly ties the user to their profile. Through the benefits of this anonymity, radicalizers are not forced to operate with the caution exercised in real-world interpersonal recruitment. This anonymous quality of social networks has “contributed to enabling immediate communication of decentralized network of terrorism”[96]. Enver Buçaj, a professor at the University of Prizren in Kosovo, contends this can be combatted through filtering and moderating any content containing terrorist persuasion in the slightest[96]. However, social media platforms have reluctance as stricter policy would negatively affect user engagement levels, attributed to the heightened interaction levels with misinformation and extreme posts[6]. Higher engagement levels drive enhanced cost of advertisement space, which affects revenue generation[29]. If this content was filtered in the suggested way, social media companies would lose a vast portion of revenue. This combination of mechanisms caters directly to the successful recruitment of citizens: persuaded to join a particular cause and the extremist community as a whole. The above analysis provides six main flaws found on social media platforms: engagement, personalization, content moderation, anonymity, machine learning models, and misinformation over truth.

Structural Weaknesses: 1) engagement priority 2) personalization 3) content moderation 4) anonymity 5) Rate of misinformation spread 6) machine learning models

Election Interfering

Throughout the history of 20th and 21st-century nations, no two countries interfered in elections as frequently or successfully as Russia and the United States. It followed that the soundest method to identify structural flaws in social platforms being abused by election interferers was an election that contained both parties. The 2016 United States Presidential Election provides an ideal opportunity to assess weaknesses.

In the years preceding and following 2016, an operation run by the General Staff of the Armed Forces of the Russian Federation (GRU), sought to interfere with the US presidential election and sow seeds of conflict. The expressed intention was to polarize liberals farther left and conservatives farther right. Through further polarization, the United States subjects itself to heightened levels of internal conflict and weakens itself. The dissension evolved into political turmoil, civil war, or similar states of agitation and lack of compromise. The main avenues conveying their persuasion messages were Facebook and Twitter, with minor activity identified on Reddit[61]. Facebook and Twitter provide numerous pieces of literature detailing the interference operation. These two platforms are the focal point of this analysis which will detail strategies, statistics, and impact.

A portion of this election interference campaign was conducted similarly to the trolling methodology detailed earlier. When analyzing Twitter activity and user logs surrounding the 2016 US presidential election, massive troll farms armed with bots, or automated accounts, can be found. In fact, automated accounts are calculated to have made up 18% of all tweets concerning the 2016 US Presidential election[85]. Russia's Internet Research Agency(IRA) was tied to this significant growth of bot accounts on Twitter[61]. Russia controlled a substantial part of the online Twitter dialogue surrounding the US presidential election with fake accounts. The core issue stemmed from the ease of account creation, which allowed their automated accounts to thrive undetected. Their anonymity remained unchallenged and free from scrutiny. The IRA operated discreetly and anonymously. Their profiles proved tricky to moderate as bios showcased a carefully selected distribution of popular topics and qualities of legitimate users: "love", "life", "trump", "conservative", "journalist", "independent", "news", "proud"[61]. This subtle phrasing allowed for their induction into online conservative groups and communities. Once inside their influence was evident by heightened interactivity of conservatives with posts versus liberals[56]. Interestingly, the conservatives documented produced 36 times more tweets than their liberal counterparts. A portion of this imbalance is attributed to the IRA campaign. Overall, this impact can be traced to the extent of personalization in the digital environment: the users felt these accounts and content resonated with their own beliefs. Despite the accounts being operated by Russian intelligence personnel thousands of miles away, US citizens social feeds started to become echo chambers of Putin's pedaled propaganda. To further this echo-chamber effect, the Twitter bot accounts amplified any content generated by real IRA operative accounts[33]. This added legitimacy awarded to posts ensured that they would be seen. Taking advantage of social media users', and humans in general, inclination to believe what is popular over what is true, or fact. Leading to messages containing hate speech and divisive opinions being spread to advance the Russian agenda of dissension. The GRU gained immense power through their expertise in exploiting Twitter, as well as Facebook, which was subject to auxiliary operations happening in parallel.

Facebook's role in the 2016 election extends beyond the GRU exploitation with persuasive messaging. Another player of interest is Cambridge Analytica, which provided confirmed aid to the Trump campaign and allegedly assisted the Russians[38]. Cambridge Analytica's power came from the attention placed on a user's activity in connection to their OCEAN analysis test. The OCEAN analysis test claims that every personality, no matter the complexity, stems from five traits. These traits are "openness, conscientiousness, extraversion, agreeableness, and neuroticism"[38]. Differences lie in the percentage a person exhibits each trait, which results in an overall assessment. Countless people pioneered this method but American psychologist Lewis Goldberg receives the majority of the credit for determining these five personality traits[111]. Cambridge Analytica offered the test to millions of Facebook users and compiled the results. Then, they established a model for determining a user's OCEAN profile from their likes, shares, comments, and other metrics on social networking platforms[38]. In this manner targeting specific individuals was perfected without needing every person to take the OCEAN test. Further, Cambridge Analytica advertised and bragged they had "5,000+ data points on 230 million US adults"[92]. Leading up to the election, the Trump campaign assigned them to influence

American citizens' voting choices in key battleground states. Persuasive technology techniques employed in real-time to craft what content should be shown to alter a voter's preference and political efficacy[38]. This sharing of data by Facebook demonstrated a lack of care and blatant disregard for users' safety and well-being. Either a preference for engagement with advertisements became indisputable or Facebook severely underestimated captology's application to its platform. Proof supports the former when analyzing Russia's Internet Research Agency's Facebook campaign.

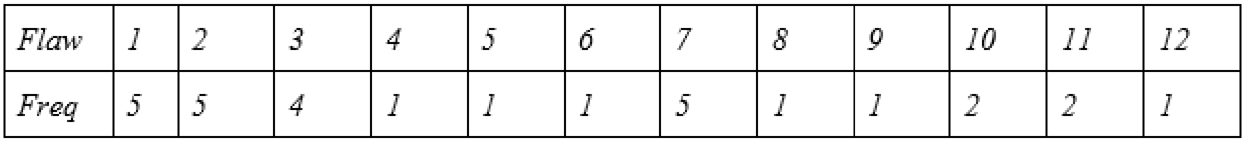
Political advertisements existed long before the internet. In the same way, nations influenced elections in other nations since the dawn of elected officials. As such certain safety measures guaranteed no country could impose undue influence undetected. The Foreign Agent Registration Act(FARA) of 1983 states the necessity for foreign actors to register with the federal government when conducting lobbying or electioneering in another nation[33]. Its application mainly resides in reducing the anonymous system that allows for successful election interference campaigns. Digital realms are a tougher place to enforce these requirements due to Virtual Private Networks, VPNs, and other identity masking tools. Due to this uncertainty and obfuscation, Facebook's liability becomes debatable. The platform witnessed behavior by the IRA specifically around advertisements. By estimates, the IRA allocated over $100,000 to targeted advertisement spending, which converted into approximately 3,500 advertisements[33]. No claims by Facebook confirm any of these advertisements were blocked. Supposedly, a stance and sentiment enabled and emboldened by legal immunity under Section 230(c) of the Communications Decency Act of 1996[34][35]. Despite there not being a single company policy these actions violated, the purchase of this space by foreign actors should have been reported and dealt with accordingly[93]. This requirement stems from the FARA of 1983. Especially, as the majority of these advertisement purchases were traceable to Russia. Again a blatant display of negligence, or ignorance as claimed by Facebook, is present. Showcasing that content moderation and election integrity equate to a diminutive concern in the company's eyes. Facebook's failure displays a wider, fundamental issue of value placed on advertisers and engagement, that content moderation severely threatens to diminish. This quantity of influence Russia gained over the American democratic process in 2016 is a minuscule illustration of the damage these structural features cause. The above case study of the 2016 US Presidential Election highlights six core flaws: personalization, anonymity, handling of data, content moderation, engagement structure, and popularity over truth.

Structural Weaknesses 1) personalization 2) anonymity 3) handling of data 4) content moderation 5) engagement structure 6) popularity over truth

Structural Weaknesses

Throughout these case studies, repeated themes of features enabled S.C.A.R.E. actors to operate efficaciously. In totality, the discussed features are the following:

1)Engagement structure 2)Content moderation 3)Anonymity 4)Invisibility 5)Non-transparency 6)Payment method 7)Personalization 8)Vitality 9)Popularity over truth 10)Data handling 11)Rate of misinformation spread 12)Machine learning models



Flaw	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Freq	*5*	*5*	*4*	*1*	*1*	*1*	*5*	*1*	*1*	*2*	*2*	*1*

Table 3. The number of S.C.A.R.E. actors abusing each feature.

They highlight the recurring structural issues of anonymity, engagement priority, content moderation, and personalization, which surfaced the most. These were the only repeated features found in every case study; therefore, the shared enablers of S.C.A.R.E.. Further, numerous features that exacerbated abuse were different expressions of the main four features. For instance 'popularity over truth' and 'rate of misinformation spread' are derived from the engagement structures of platforms. Also noted is that the anonymity feature was not identified as key to advertisers, however, it received widespread abuse by every other actor. Additionally, this supports the table depicting how the Honeycomb framework links to S.C.A.R.E. actors. The below table completes the missing features which resulted in the abuses of the social media function.

Function	**Feature**	**Abuse Type**	**S.C.A.R.E. actor**
Sharing	Content Moderation, Engagement	Inappropriate UGC	S.C.A.R.E.
Groups	Personalization	Echo-chambers/Bias	S.C.A.R.E.
Conversation	Engagement	Misinformation/Disinformation/Aggressive engagement	S.C.A.R.E.
Identity	Anonymity, Personalization	Target of exploitation(user)	S.C.-.R.E.
Relationships	S.C.A.R.E. Actors	Threat/Coercion/Abuse/Intimidation	S.C.A.R.E.
Reputation	NA	Shaming/Defamation	-.C.-.-.-.
Presence	NA	Location tracking/Monitoring	-.-.-.-.-.

Table 4.The Honeycomb framework and S.C.A.R.E. actor analysis with features

Now it becomes possible to view each function as being implemented as a feature manifesting into abuse. Engagement and personalization are the two features coming from more than one function and leading to two outcomes of abuse. The importance of these features informs how to weigh the evaluation metrics in the methodology section.

The identified features compared to another study's findings serves to add legitimacy to the assessment. For this, a report by the Omidyar group contends that six features of social media platforms are threatening our democracy as a whole[85]. This cross-referencing of the factors found in the case studies with this list can weigh the validity of features contributing to S.C.A.R.E. exploitation. Each issue discussed in the report linked to a structural weakness provides preliminary support. Then excerpts and definitions corroborate the assignments. Subsequently, the synthesis addresses which features to use in the methodology section and those to ignore. A complete overlap proves substantive to proceed in setting up an evaluation metric with the four features as elements.

1. "Echo chambers, polarization, and hyper-partisanship."[85]

    (1) Personalization

    Ø *Echo chambers:* "The prioritization of user preferences results in a feedback loop where the feeding of news, search results, and social network updates that align with user attitudes and interests exacerbates and reinforces user preferences"[85]. Previously defined as being an element of personalization.

    Ø *Hyperpatisanship*: A term indicative of identity politics where an individual is summarized by the political ideology they hold[85].

2. "Manipulation, micro-targeting, and behavior change."[85]

(1) Personalization

Ø *Microtargeting:* “To direct tailored advertisements, political messages, etc., at (people) based on detailed information about them (such as what they buy, watch, or respond to on a website)” [86]. A captology technique present with actors such as Cambridge Analytica to increase user interaction with content.

Ø *Behavior change:* The report discusses the role user data plays in changing behavior, which is characteristic of personalization techniques.

3. “Political capture of platforms.”[85]

(1) Engagement priority

Ø *Political capture:* A study by Stanford University’s Business school reveals the ability of advertisements to shift election results and the inherent role they play in politics[98]. Drawing a relationship between politics and advertisements suggests advertisement dominance of a platform is political dominance. Further, the priority of engagement platforms stems from advertiser influence over features as demonstrated by the Facebook case study.

4. “Proliferation of several types of misinformation and disinformation.”[85]

(1) Engagement priority

Ø *Proliferation: “*Social network platforms have huge incentives to accommodate the creation and distribution of content and feed the "attention economy.""[85] The rapid spread of information and engagement by users is the result of advertisement influence on priorities.

(2) Content moderation

Ø *Misinformation and disinformation:* This highlights a moderation issue of failing to identify this damaging content.

5. "Intolerance, exclusion of disadvantaged or marginalized voices, public humiliation, and hate speech"[85]

(1) Content moderation

Ø *Intolerance, exclusion, public humiliation, hate speech:* Behavior exhibited in chaos causers, or trolls, which stems from poor content moderation practices of platforms.

6. "Conflating popularity, legitimacy, and user intentionality."[85]

(1) Anonymity

Ø *Conflating popularity, legitimacy:* The report specifies the rise of unknown individuals challenging the statements and rulings of experts that are verified for authenticity. Further, displaying the power that has been given to the anonymous user.

Chapter 4

Methodology

The process and outline followed in the Methodology and Data sections build on the four structural weaknesses allowing for S.C.A.R.E. actors to abuse users. These structural flaws include content moderation, anonymity, engagement priority, and personalization.

Process

For this research, a metric is created for each of the four common structural issues identified in the case studies section. The identified weakness receives a percentage rating based on how present it is on the platform. Then, an overall percentage indicating the propensity to be abused by S.C.A.R.E. actors will be tabulated. Finally, the ratings are normalized and compared for analysis.

Data Sources

These equations are severely limited by the amount of transparency social networking platforms have with the public. Thus it follows that the more tremendous pressure the public places on companies to report factual data, the more accurate these metrics can be. Data for metric computation will be derived from a variety of accredited sources. Due to social media platforms' close hold on statistics and information, reliance on third-party assessors and statistic firms forms. When possible, transparency reports and official releases will inform the data. Slight approximations may deem necessary to get figures that align with the formulas as no formal statistic exists online. Furthermore, the gathered data ranges from 2018 to 2022 because of availability and sporadic reporting.

Selected Platforms Rationale

The four social media platforms of Facebook, YouTube, Reddit, and Twitter were selected for evaluation. These platforms were selected for two reasons. First, they were the most observed platforms when conducting research for the case studies component. Second, the United States January 6th panel responsible for investigating the riot's circumstances after the 2020 US presidential election requested fifteen companies to turn over 'misinformation' and 'extremism' records[32]. Four companies ignored or refused these requests and were subsequently issued subpoenas. The social media companies included Alphabet(parent company of YouTube), Meta(parent company of Facebook), Reddit, and Twitter. Therefore, a combination of popularity, the frequency of abuse, and failure to turn over records deem these platforms worth investigating and evaluating.

Anonymity Metric

The ability for users to operate anonymously on social media platforms originates from the personal information required for account creation. This information includes various combinations of inputs such as name, email, phone number, password, birthday, and/or gender. The fewer of these required, the easier to make fake accounts. Prior elaboration on bot accounts and troll farms emphasizes the need for comprehensive account signups. The sophistication of verification methods, such as the Computer Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA), adversely affects users' creation of fake accounts. Additionally, platforms implementing email, or phone, verification messages after an account is created demonstrate a heightened level of complexity and security. The extent is quantitatively measured as the duration of a given signup process, which affects the number of accounts created in a timeframe. Some platforms offer alternative options for account creation by linking Facebook, Google, Apple, or other accounts. This lessens the signup complexity due to offloading trust and user verification steps to third-party platforms, thereby improving the ease of fake account creation. The proven method requiring a photo ID to be uploaded is seen on extremely secure platforms, primarily when related to finance applications[51]. A rating scale devised from these factors yields the following system where a point is added for every feature missing in registration. Then, the equation below solves for the percentage of missing features.

Six login features:

1. No CAPTCHA
   a. A CAPTCHA prevents bots or scripts from creating accounts. If one is present, the platform receives a 0 for this feature.
2. No email or Phone verification message
   a. Limiting one account per phone number and/or email slows the adversary down. This adds more steps necessary to make fake accounts, therefore a platform with a verification system receives a 0.
3. Less than 1 minute to register

$$330,000,000*0.035=11,550,000 \text{ accounts}$$

   a. The United States population is roughly 330,000,000 people. The figure 11,550,000 people is reached through the application of the 3.5% rule. This

3.5% number represents the minimum threshold of participation in which a successful revolution is guaranteed[113]. It is the percentage of a population that any political movement needs to enlist to make effective change. If no automated programs are factored in, a group of 20 people could create this many accounts in a year. This is contingent upon signup time being less than a minute. Therefore, this metric theoretically determines how easily a platform could be utilized to influence a country.

4. Third-party account creation
    a. Third-party sign in options offload trust thus weakening any sign-in security.
5. Does not require all the following: name, email, phone number, password, birthday
    a. The more information needed from a user, the harder it is to create an anonymous account. Moreover, the ability to identify fakes improves.
6. No photo ID required
    a. Photo ID gives a likely assurance the person is legitimate.

TP: *total points of account creation features*

AS: *anonymity score as a percentage*

$$AS = TP/6$$

Content Moderation Metric

Definitive guidelines for content moderation support secure online environments. Platforms vary vastly in procedures and policies surrounding content moderation; therefore, a statistical evaluation determines the percent of moderated content. Despite transparency pages and reports becoming common on these platforms circa 2019, it proved difficult to assess the real numbers[44]. The information detailed in these transparency reports tends to showcase social media platforms in a positive light with unbelievably successful moderation policies. Hopefully, the passage of time forces increasingly accurate reports. Consequently, a new formula grounded in data from transparency reports and baselines from trustworthy independent parties deems itself necessary.

B: *the percent of the total content that should be monitored; this is a constant*

MC: *the amount of content moderated as reported by transparency reports*

IMC: *the amount of content incorrectly moderated and reversed*

TC: *the total amount of content posted*

MS: *moderation score as a percentage; harmful content left unmoderated*

$$MS = 100\% - (MC - IMC)/(B * TC)$$

The equation seeks to determine the percent of harmful content that platforms neglect or miss when moderating. The count of moderated content excludes incorrectly labeled or moderated content. When verifiable estimates of the percent of harmful content do not exist, a baseline percentage of moderated content substitutes. It is near impossible to calculate how much content is not moderated as the social networking platforms, specifically, Facebook, has penetrated a quarter of the world population[57]. Therefore, Facebook’s internal research into the issue provides a constant variable. While this baseline is not ideal, it is the best available method to calculate moderation efforts. In 2021, the Facebook whistleblower, Frances Haugen, exposed that based on internal calculations the platform removes “3%-5% of hateful content” and “0.6% of content invoking violence or incitement” [44]. Considering these figures and assuming the best-case scenario for Facebook, we approximate 5% as the percentage of all harmful content that undergoes moderation. The number will be deducted from 100% to show the moderation failure of Facebook. In the data component, this number replaces **MS**, the amount of harmful content left unmoderated and assigns a value for **B**, the baseline of all content on social media that should be moderated. Admittedly, a key assumption that all platforms approximately share this percentage exists.

Engagement Priority Metric

The degree to which a company serves its users in relation to third parties determines the extent of a platform’s engagement priority. This study assesses the tradeoff by the proportion of a company’s total revenue generated from advertisers' contributions. This is a litmus test on how much consideration from advertiser stakeholders informs a platform’s choices and priorities. For instance, a company receiving 0% of revenue from advertisements does not consider advertiser needs. In contrast, a company receiving 100% of revenue from advertisements must adopt the needs and goals of advertisers as their own. As discussed in the previous section, advertisers rely on users buying products or changing beliefs directly related to social media platforms’ implementation of their advertisements. Therefore, a cause and effect relationships forms. The greater the engagement levels are, the more revenue contributed by advertisers. With this framing, the percent of the revenue generated from advertisements is construed as the magnitude of priority placed on engagement. This evaluation metric for engagement priority calculated by the following formula.

TR: *total revenue generated*

AR: *revenue generated from advertisements*

ATR: *percent of total revenue generated from advertisements*

EPS: *engagement priority score as a percentage*

$$EPS = ATR = AR/TR$$

Personalization Metric

The extent of personalization is calculated as a function of user interaction and time spent on the application. Successful personalization can properly produce content that the user will choose to interact with on a regular basis. Moreover, user interaction is understood as a function of the degree to which a platform is optimized for personalization. Noticeable impacts of success are echo-chamber environments with users that share eerily similar beliefs. Under this premise, the metric can be devised as a formula comparing the platform's average click-through rate to the industry standard. This figure is then averaged with the percentage of users who go on the platform daily, which indicates a habit formed from personalization[5]. The variables are utilized to create a personalization metric considering the three variables defined below: BCTR, CTR, and DUR. A baseline click-through rate of 0.35% is chosen as it is the average rate across all industries and platforms[89]. If a CTR is below 0.35%, it will be replaced by the baseline number, because a score less than the baseline represents a negligible level of personalization. The average CTR per platform references independent analyses due to the difficulty of accessing such figures. For context, the following formula is the accepted standard for calculating click-through rates[89]. The subsequent formula is the evaluation metric for personalization.

$$CTR = (clicks/impressions) * \mathbf{100\%}$$

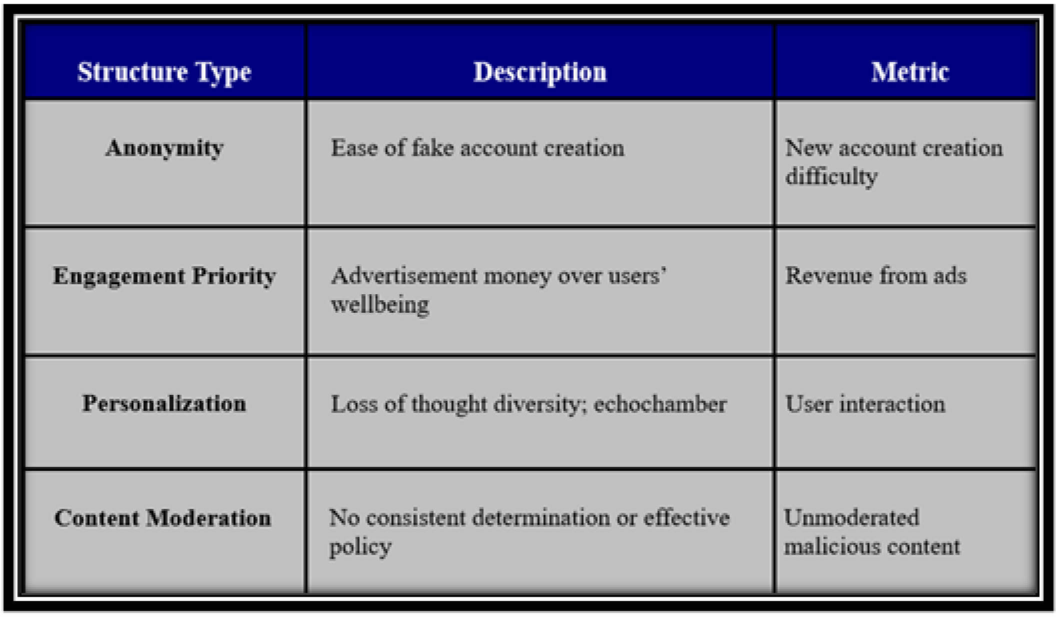
BCTR: *constant click-through rate; 0.35%*

DUR: *percent of users who use the platform daily*

CTR: *average click-through rate of platform*

PS: *personalization score as a percentage*

$$PS = ((100\% - BCTR/CTR) + DUR)/2$$



Structure Type	Description	Metric
Anonymity	Ease of fake account creation	New account creation difficulty
Engagement Priority	Advertisement money over users' wellbeing	Revenue from ads
Personalization	Loss of thought diversity; echochamber	User interaction
Content Moderation	No consistent determination or effective policy	Unmoderated malicious content

Table 5. Features with description and metric

Calculation of Scores

The analysis completed in the case studies section and subsequent combination of features with honeycomb functions and abuses determines how to weigh each contributor in the overall S.C.A.R.E. score. Personalization and engagement are the two features that stem from more than one function and lead to two types of abuse. Therefore, these two features will be given a 1.5x multiplier in the overall calculation. This multiplier is chosen to give both of these features an increased pull in the overall rating, but counting each as if they were two features disproportionately slants the results to be dependent only on these two features. Furthermore, assigning 50% increases brings the overall weight to five: the number of S.C.A.R.E. actors.

SS: S.C.A.R.E. score

$$SS = (1.5)PS + (1.5)EPS + MS + AS$$

Normalization of Scores

Based on each score given to a platform, a normalization process is undergone to compare platforms concerning the number of users. This stems from the contention that the more popular an item is, the greater the risk of misuse. The global penetration rate is factored in, which is the number of active users on the platform compared to all users active on social media. The NS normalization figure has a weight of one, thereby bringing the total weight of features to six.

AUP: *active users on the platform*

AUSM: *active users on social media*

NS: *percent of all active users*

NSS: *normalized S.C.A.R.E. score*

$$NS = AUP/AUSM$$

$$NSS = (1.5)PS + (1.5)EPS + MS + AS + NS$$

Chapter 5

Data

Assessing Platforms

Facebook

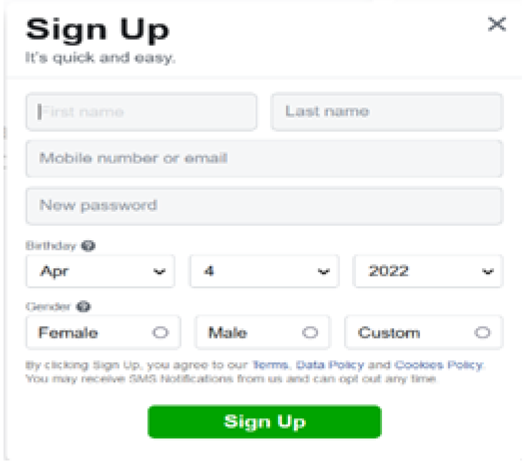
Anonymity Score

Score: 50%, 3/6

Support: [41]

Point	Explanation of Login Feature
0	*CAPTCHA present*
0	*Email or Phone verification message*
1	*Less than 1 minute to register*
0	*No third-party account creation*
1	*Does not require all of the following: name, email, phone number, password, birthday*
1	*No photo ID required*
3	**Total out of 6**

Table 6. Facebook Anonymity Rating



Sign Up

It's quick and easy.

First name | Last name

Mobile number or email

New password

Birthday

Apr | 4 | 2022

Gender

Female ○ | Male ○ | Custom ○

By clicking Sign Up, you agree to our Terms, Data Policy and Cookies Policy. You may receive SMS Notifications from us and can opt out any time.

Sign Up

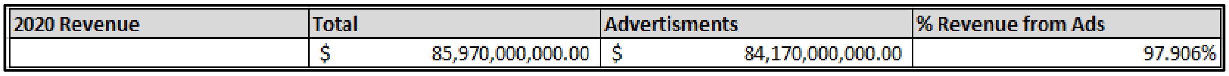
Figure 10. Facebook Signup Page

Figure 9. Facebook login[41].

Engagement Priority Score

Score: 97.9%

Support: [54][55]



2020 Revenue	Total	Advertisments	% Revenue from Ads
	$ 85,970,000,000.00	$ 84,170,000,000.00	97.906%

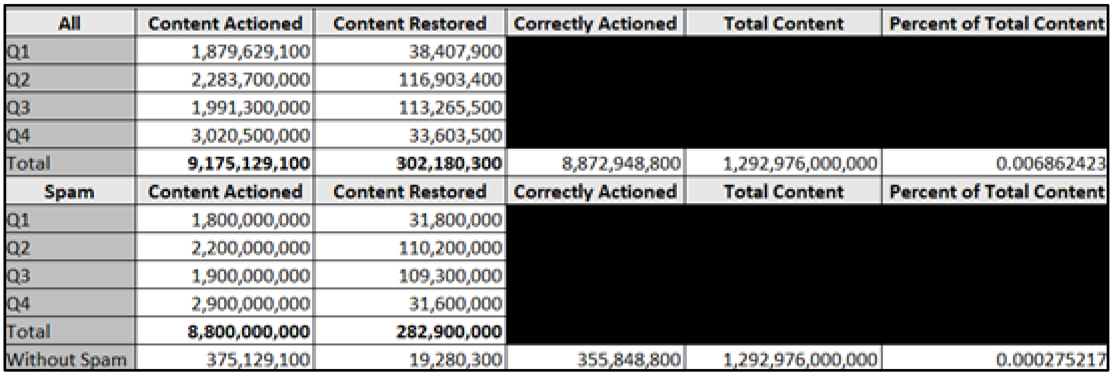
Table 7. Facebook EPS

Content Moderation Score

Score: 95%

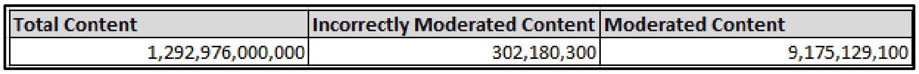
Support: [42] [40]

$$MS = 100\% - (9{,}175{,}129{,}100 - 302{,}180{,}300)/(0.1372 * 1{,}292{,}976{,}000{,}000)$$



All	Content Actioned	Content Restored	Correctly Actioned	Total Content	Percent of Total Content
Q1	1,879,629,100	38,407,900	[redacted]	[redacted]	[redacted]
Q2	2,283,700,000	116,903,400	[redacted]	[redacted]	[redacted]
Q3	1,991,300,000	113,265,500	[redacted]	[redacted]	[redacted]
Q4	3,020,500,000	33,603,500	[redacted]	[redacted]	[redacted]
Total	**9,175,129,100**	**302,180,300**	8,872,948,800	1,292,976,000,000	0.006862423
Spam	**Content Actioned**	**Content Restored**	**Correctly Actioned**	**Total Content**	**Percent of Total Content**
Q1	1,800,000,000	31,800,000	[redacted]	[redacted]	[redacted]
Q2	2,200,000,000	110,200,000	[redacted]	[redacted]	[redacted]
Q3	1,900,000,000	109,300,000	[redacted]	[redacted]	[redacted]
Q4	2,900,000,000	31,600,000	[redacted]	[redacted]	[redacted]
Total	**8,800,000,000**	**282,900,000**	[redacted]	[redacted]	[redacted]
Without Spam	375,129,100	19,280,300	355,848,800	1,292,976,000,000	0.000275217

Table 8. Facebook CMS[40].



Total Content	Incorrectly Moderated Content	Moderated Content
1,292,976,000,000	302,180,300	9,175,129,100

Table 9. Facebook CMS Calculation

Calculation of B constant

To solve for B, where 95% of harmful content is not moderated:

$$95\% = 100\% - 0.686\% * B$$

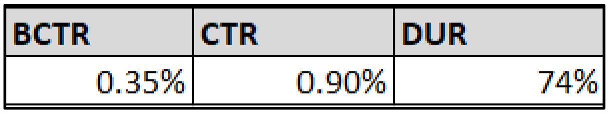
B: 13.72% of all content should be moderated

Personalization Score

Score: 67.5%

Support: [52] [45]

$$PS = ((100\% - (0.35\%/0.89\%)) + 74\%)/2$$



BCTR	CTR	DUR
0.35%	0.90%	74%

Table 10. Facebook PS

Twitter

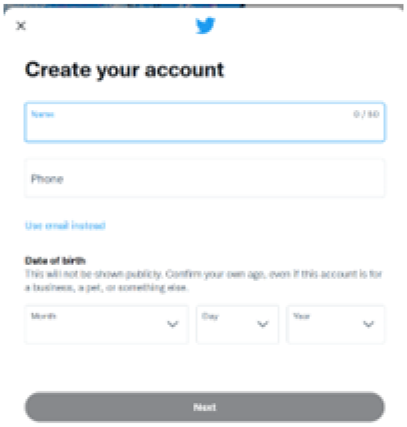
Anonymity Score

Score: 83.3%, 5/6

Support: [43]

Point	Explanation of Login Feature
1	No CAPTCHA present
0	Email or Phone verification message
1	Less than 1 minute to register
1	Third-party account creation (Apple, Google)
1	Does not require all of the following: name, email, phone number, password, birthday
1	No photo ID required
5	**Total out of 6**

Table 11. Twitter Anonymity Rating



Create your account

Date of birth

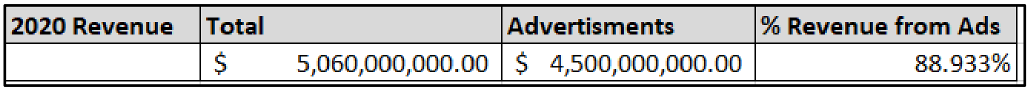
This will not be shown publicly. Confirm your own age, even if this account is for a business, a pet, or something else.

Figure 11. Twitter login[43]

Engagement Priority Score

Score: 88.9%

Support: [46]



2020 Revenue	Total	Advertisments	% Revenue from Ads
	$ 5,060,000,000.00	$ 4,500,000,000.00	88.933%

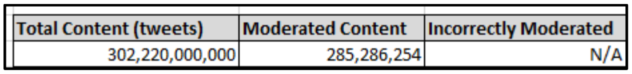
Table 12. Twitter EPS

Content Moderation

Score: *99.3%*

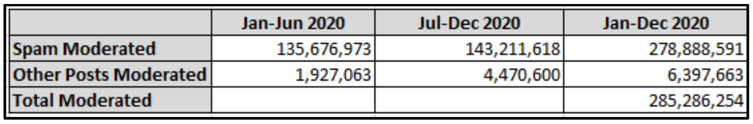
Support: *[100][47]*

$$MS = 100\% - (285{,}286{,}254 - 0)/(0.1372 * 302{,}220{,}000{,}000)$$



Total Content (tweets)	Moderated Content	Incorrectly Moderated
302,220,000,000	285,286,254	N/A

Table 13. Twitter CMS Calculation



	Jan-Jun 2020	Jul-Dec 2020	Jan-Dec 2020
Spam Moderated	135,676,973	143,211,618	278,888,591
Other Posts Moderated	1,927,063	4,470,600	6,397,663
Total Moderated			285,286,254

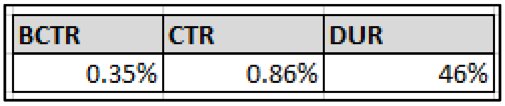
Table 14. Twitter CMS

Personalization Score

Score: 52.7%

Support:[52] [48]

$$PS = ((100\% - (0.35\%/0.86\%) + 46\%)/2$$



BCTR	CTR	DUR
0.35%	0.86%	46%

Table 15. Twitter PS

Youtube

Anonymity

Score: 66.7%, 4/6

Support: [99]

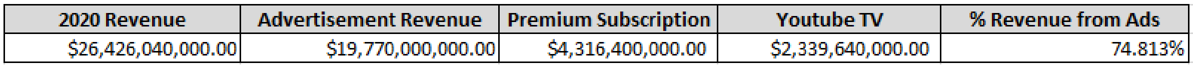
Point	Explanation of Login Feature
1	No CAPTCHA present
0	Email or Phone verification message
1	Less than 1 minute to register
0	No third-party account creation
1	Does not require all of the following: name, email, phone number, password, birthday
1	No photo ID required
4	**Total out of 6**

Table 16. YouTube Anonymity Rating

Engagement Priority Score

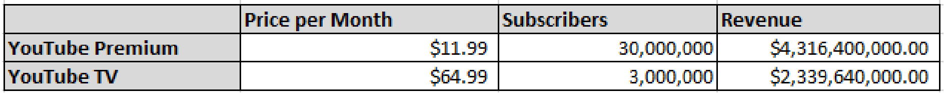
Score: 74.8%

Support: [49]



2020 Revenue	Advertisement Revenue	Premium Subscription	Youtube TV	% Revenue from Ads
$26,426,040,000.00	$19,770,000,000.00	$4,316,400,000.00	$2,339,640,000.00	74.813%

Table 17. YouTube EPS Calculation



	Price per Month	Subscribers	Revenue
YouTube Premium	$11.99	30,000,000	$4,316,400,000.00
YouTube TV	$64.99	3,000,000	$2,339,640,000.00

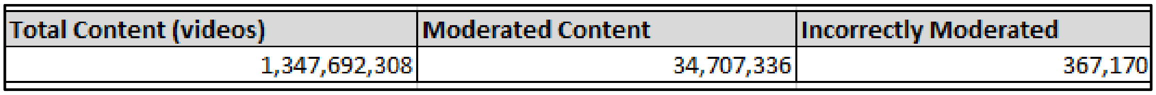
Table 18. YouTube EPS

Content Moderation Score

Score: 81.4%

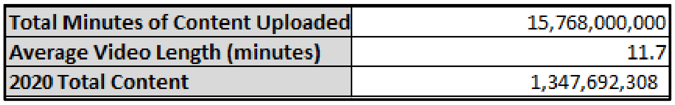
Support: [90] [91] [101]

$$MS = 100\% - (34{,}707{,}336 - 367{,}170)/(0.1372 * 1{,}347{,}692{,}308)$$



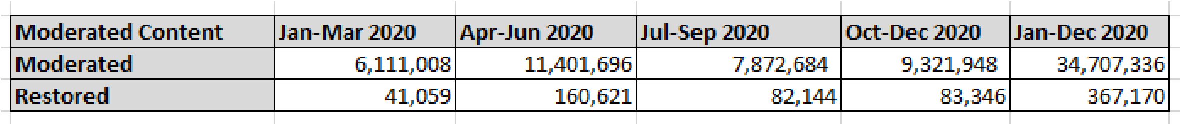
Total Content (videos)	Moderated Content	Incorrectly Moderated
1,347,692,308	34,707,336	367,170

Table 19. YouTube CMS Calculation



Total Minutes of Content Uploaded	15,768,000,000
Average Video Length (minutes)	11.7
2020 Total Content	1,347,692,308

Table 20. YouTube Total Content



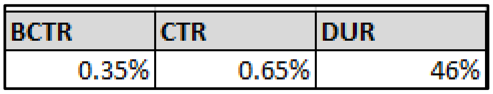
Moderated Content	Jan-Mar 2020	Apr-Jun 2020	Jul-Sep 2020	Oct-Dec 2020	Jan-Dec 2020
Moderated	6,111,008	11,401,696	7,872,684	9,321,948	34,707,336
Restored	41,059	160,621	82,144	83,346	367,170

Table 21. YouTube Moderation

Personalization Score

Score: 46.1%

Support: [52][50]



BCTR	CTR	DUR
0.35%	0.65%	46%

Table 22. YouTube PS

$$PS = (100\% - (0.35\%/0.65\%) + 46\%)/2$$

Reddit

Anonymity

Score: 83.3%, 5/6

Support:[58]

Point	Explanation of Login Feature
0	CAPTCHA present
1	No Email or Phone verification message
1	Less than 1 minute to register
1	Third-party account creation(Apple,Google)
1	Does not require all of the following: name, email, phone number, password, birthday
1	No photo ID required
5	**Total out of 6**

Table 23. Reddit Anonymity Rating

Engagement Priority Score

Score: 100%

Support: [59][60]

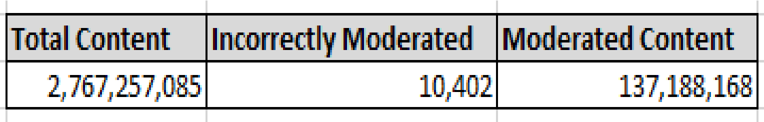
Reddit offers a premium account subscription on the platform, which allows a user to pay to get rid of advertisements. However, a recent revamp of the platform could see this disappear. Reddit only recorded advertisement revenue for the past years and no information is given on the number of premium users. There was $181.3 million in revenue recorded for 2020[59]. With this lack of data and transparency, Reddit is credited with having its entire revenue from advertisements.

Content Moderation Score

Score: 63.9%

Support: [79]

$$MS\ =\ 100\%\ -\ (137{,}188{,}168 - 10{,}402)/(0.1372 * 2{,}767{,}257{,}085)$$



Total Content	Incorrectly Moderated	Moderated Content
2,767,257,085	10,402	137,188,168

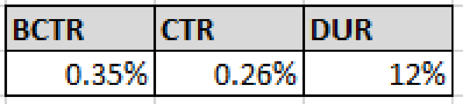
Table 24. Reddit CMS

Personalization Score

Score: 6%

Support: [87][60]

$$PS\ = ((100\% - 0.35\%/0.35\%) + 12\%)/2$$



BCTR	CTR	DUR
0.35%	0.26%	12%

Table 25. Reddit PS

Calculation of Scores

S.C.A.R.E. Scores

	Facebook	Twitter	YouTube	Reddit
Anonymity	*50%*	*83.30%*	*66.70%*	*83.30%*
Engagement Priority	*97.90%*	*88.90%*	*74.80%*	*100%*
Content Moderation	*95%*	*99.30%*	*81.40%*	*63.90%*
Personalization	*67.50%*	*52.70%*	*46.10%*	*6%*
S.C.A.R.E. Score	78.60%	79%	65.90%	61.20%

Table 26. S.C.A.R.E. Scores

$$SS = (1.5)PS + (1.5)EPS + MS + AS$$

Penetration Rate by Platform

Platform	Active users	% of all active users
Facebook	*2,797,000,000*	*64.60%*
Twitter	*396,000,000*	*9.15%*
YouTube	*2,291,000,000*	*52.91%*
Reddit	*430,000,000*	*9.99%*
All	*4,330,000,000*	*100%*

Table 27. Social Media Penetration Rates[88]

Normalized S.C.A.R.E. Scores

	Facebook	Twitter	YouTube	Reddit
Anonymity	*50%*	*83.30%*	*66.70%*	*83.30%*
Engagement Priority	*97.90%*	*88.90%*	*74.80%*	*100%*
Content Moderation	*95%*	*99.30%*	*81.40%*	*63.90%*
Personalization	*67.50%*	*52.70%*	*46.10%*	*6%*
Personalization	*64.60%*	*9.10%*	*52.90%*	*10%*
S.C.A.R.E. Score	76.30%	67.40%	63.70%	52.70%

Table 28. Normalized S.C.A.R.E. Scores

Chapter 6

Discussion and Analysis

In the previous section, the evaluation metrics indicated substantial weaknesses within the four chosen platforms. These comparative metrics resulted from the identification of four common structural flaws S.C.A.R.E. actors rely on: anonymity, engagement priority, content moderation, and personalization. This study quantifies the risk of abuse by a S.C.A.R.E. actor through a percentage score. The equations attempted to assign the percentage of presence of each flaw on the platforms Reddit, Facebook, Twitter, and YouTube, which ultimately combined into an overall rating. The higher a platform scored, the more likely the platform's abuse is. Through analysis of the calculated data, this study concludes every selected platform displays a moderate to high likelihood of abuse. This risk is a factor of the S.C.A.R.E. actors, government regulation, and platform negligence. The graph below displays the presence of each structural weakness and the overall S.C.A.R.E. score before normalization.

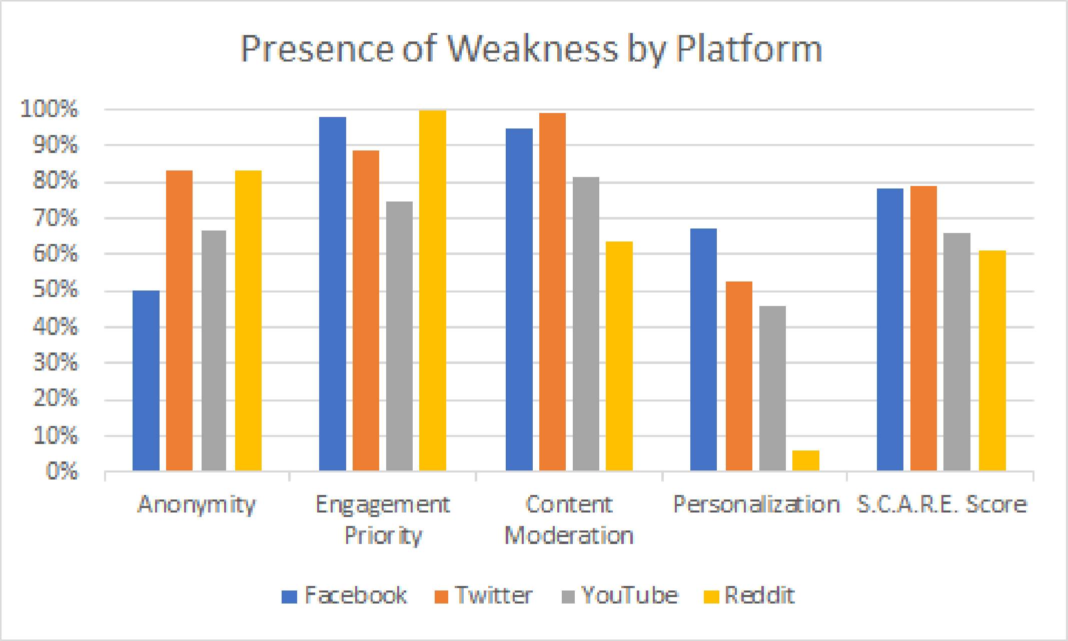


Figure 12. Structural Weakness Comparison by Platform

All four structural weaknesses identified in the case studies section demonstrated presence on Facebook, Twitter, YouTube, and Reddit. This indicates a high likelihood of abuse by S.C.A.R.E. actors. In fact, every platform scored between 60%-80% before normalization and above 50% after factoring in social media penetration rates. This existence of common structural issues that S.C.A.R.E. actors abuse across the evaluated social media platforms fulfills the first research question. Notably, the personalization evaluation metric generated significantly lower levels of presence than the rest. There are two explanations: the evaluation metric is flawed or personalization rates are lower in presence than other issues. The possibility of a flawed metric presents an opportunity for deeper understanding. Each calculation takes into account the average click-through rate of a platform's content. Essentially, how effectively can a platform cause a user to click on an advertisement? Viewed in this light, personalization results from the increased presence of an engagement priority. The magnitude of control advertisers exert over social networking platforms stokes the desire to convince users to click on advertisements. As stated previously, users interact with personalized content at a higher rate than non-personalized content. This means over time personalization levels should hypothetically equal engagement priority figures. A lack of time or a faulty metric explains the difference between the presences. In particular, Reddit shows a substantial difference between its engagement priority and personalization scores. In my research, Reddit's score inconsistency originates from a newfound emphasis on advertisements. Over the last few years starting in 2017, Reddit announced its intention to become increasingly profitable. From then until 2021, revenue tripled and investor interest soared. Investors supplied approximately $1 billion to the platform based on their user growth rate and expressed intention to expand advertisement operations[60]. Reddit's lagging personalization score is a factor of time. Given a few more years, these numbers are expected to become closer and mirror the trend exhibited by the other platforms.

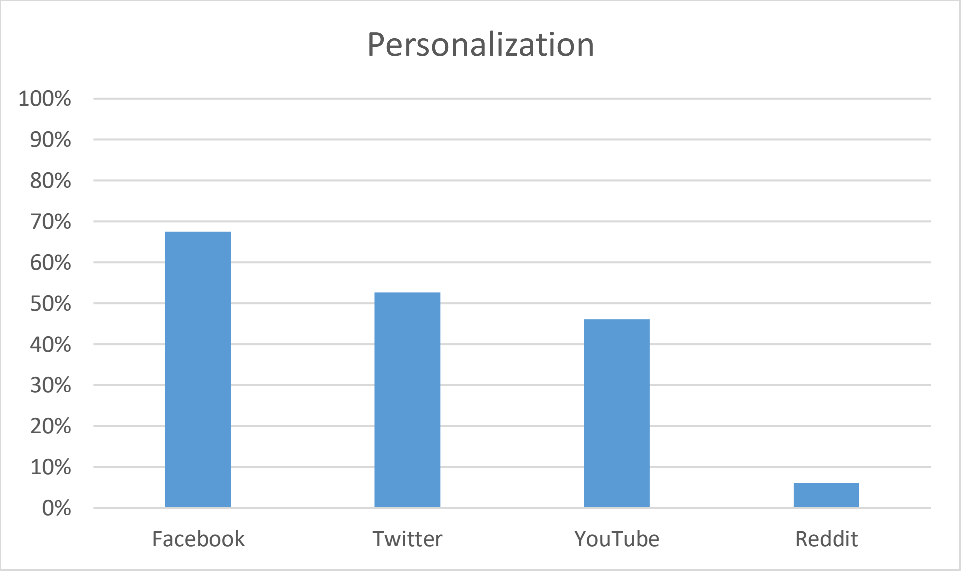


Figure 13. Personalization Scores

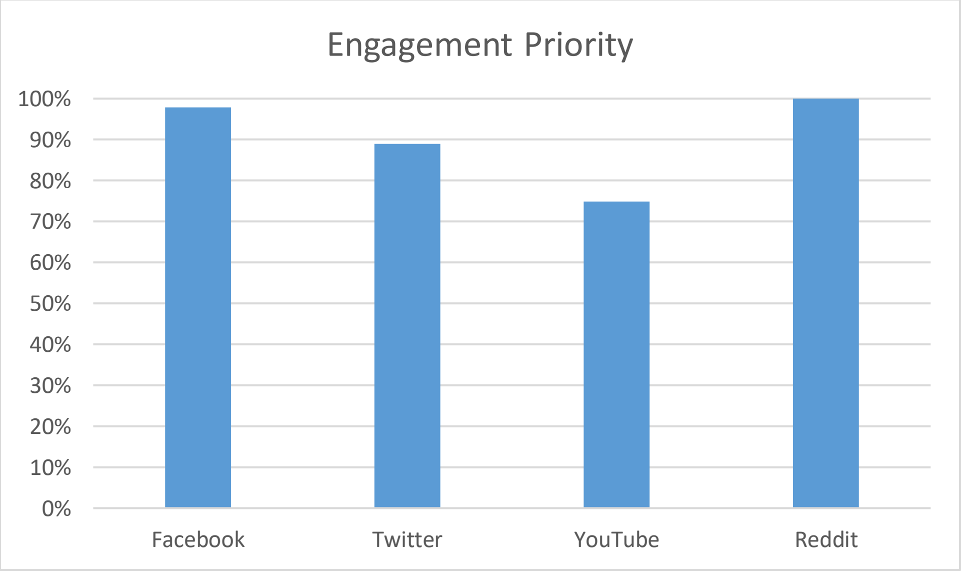


Figure 14. Engagement Priority Scores

Social media companies' priorities exposed themselves when scoring the engagement priority of each. While engagement is necessary for a platform to thrive, the eclipse of advertisers' interests over users' is not. These scores depict a depressing reality of the motives of social networking platforms. They market users to companies by employing captology principles. The high scores indicate how a platform decides whether to take steps that ensure a safe community or turn a blind eye so that advertisements may thrive. Some platforms display attempts at diversifying their revenue streams. YouTube showcases the best methods of the four. Premium accounts and YouTube TV account for nearly 30% of revenue. A prime example of profits achieved through alternative methods. Unfortunately, YouTube is the outlier among evaluated platforms and it is still above 70%. The engagement priority scores demonstrate utter dominance of profit over safety. This mindset affects how the platform conducts its responsibilities including content moderation.

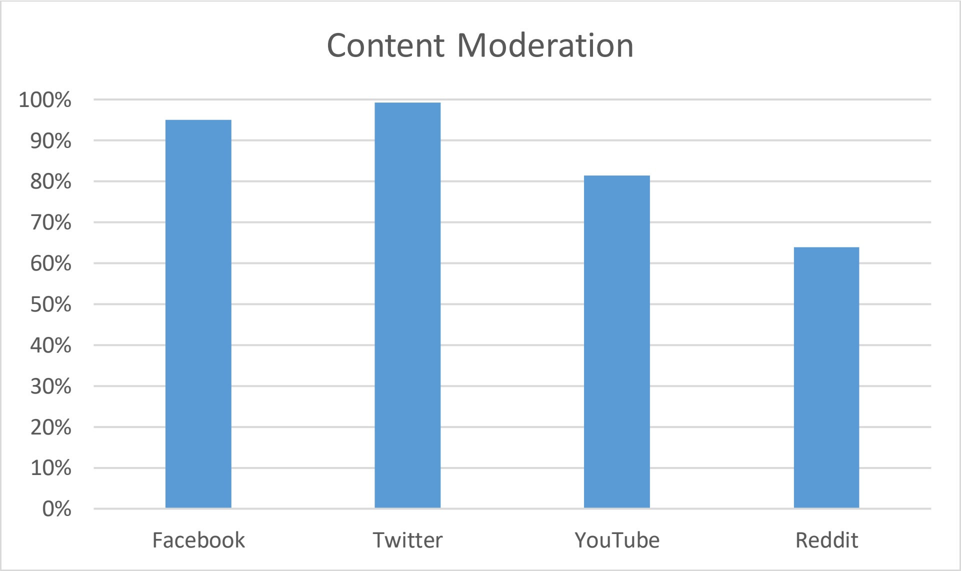


Figure 15. Content Moderation Scores

Content moderation is a notoriously tricky policy to enforce. Analyzing the content moderation scores and correlating them to content moderation systems reveals best practices in place today. Facebook and Twitter scored the worst. An expected outcome based on popular sentiment, new stories, and admittances by their own executives. Twitter's score validates its former CEO's statement that they "suck" at moderating content[83]. YouTube's rating beats the prior two, however, the rating is potentially artificially low. The future studies' areas of concern section covers the viability of the YouTube content moderation score. Finally, Reddit led the content moderation scores. Openly accessible and clear statistics in Reddit's transparency report made the calculation the easiest by far. Reddit employs a multi-faceted content moderation model: automated removals, admin removals, and moderator removals. All discussed platforms engage in the first two types of removals, but the third makes Reddit special. It enlists Reddit users to self-moderate subreddit pages and serve as their own police. A guideline provides users with black and white while letting users decide on the gray areas. Chris Slowe, the CTO of Reddit, commented on their methodology for moderation: "Our underlying approach is that we want communities to set their own cultures, policies, and philosophical systems. To make this model function, we need to provide tools and capabilities to deal with the [antisocial] minority."[115] This technique works to an extent currently. In the future, Reddit's model faces the test of time and advertiser influence. Further questions arise about how restrictive content moderation policies can be without infringing on free speech. The evaluation metric's computations correlation with platforms' policies illustrates partial success. A significant challenge to content moderation is the anonymity platforms provide. The ability for users to create fake accounts and post hateful messages from this profile demonstrates a need for secure account signups.

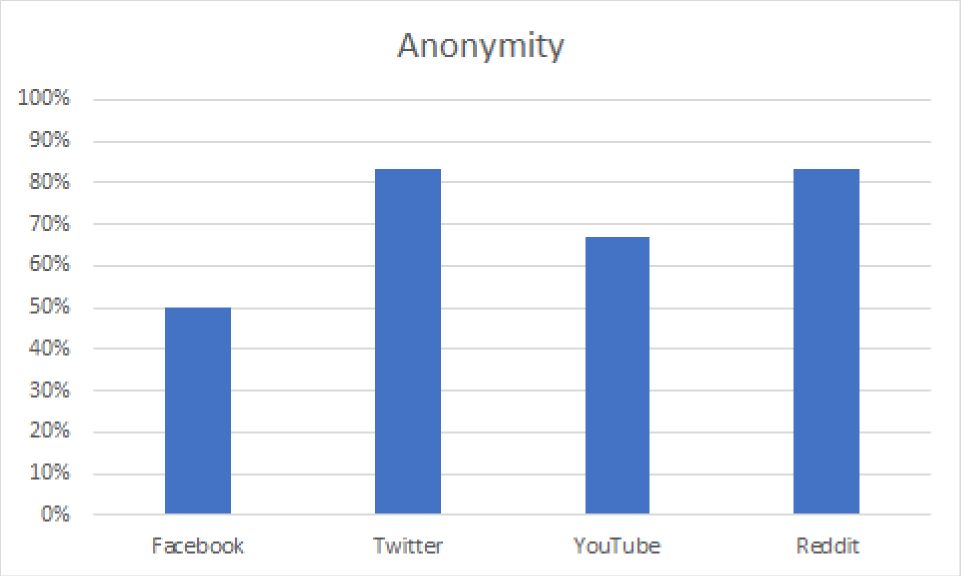


Figure 16. Anonymity Scores

The study evaluated each platform's signup process based on the presence of six components. The percentage of these six components missing represents the ease by which a S.C.A.R.E. actor creates an anonymous account. Despite high ratings in the three prior evaluations, Facebook scored the best in anonymity. In fact, it was the only platform to flag the fake account created for this study. Shortly following, it resolved the account as legitimate and successfully registered. This was the only reported difficulty with signing up a fake account on the four platforms. A likely explanation for Facebook's heightened concern for secure account creation stems from the fallout over abuses in the 2016 Presidential election and concurrent events[56]. On the other hand, Twitter and Reddit demonstrated minimal account signup complexity. The Twitter account creation process is simplistic. Automation is feasible as it lacks a CAPTCHA and allows for third-party account signup from Apple or Google. This supports the Russian IRA's choice in exploiting the Twitter platform for their 2016 operations[61]. Mainly based on the large volume of bot accounts made by the IRA. Through this analysis, it becomes evident that anonymity is possible on any platform. Moreover, all these platforms lend themselves to S.C.A.R.E. abuse. The metric detailed relies on account signup features, but there are more factors that contribute to anonymity. The future studies section explores these shortcomings in greater depth.

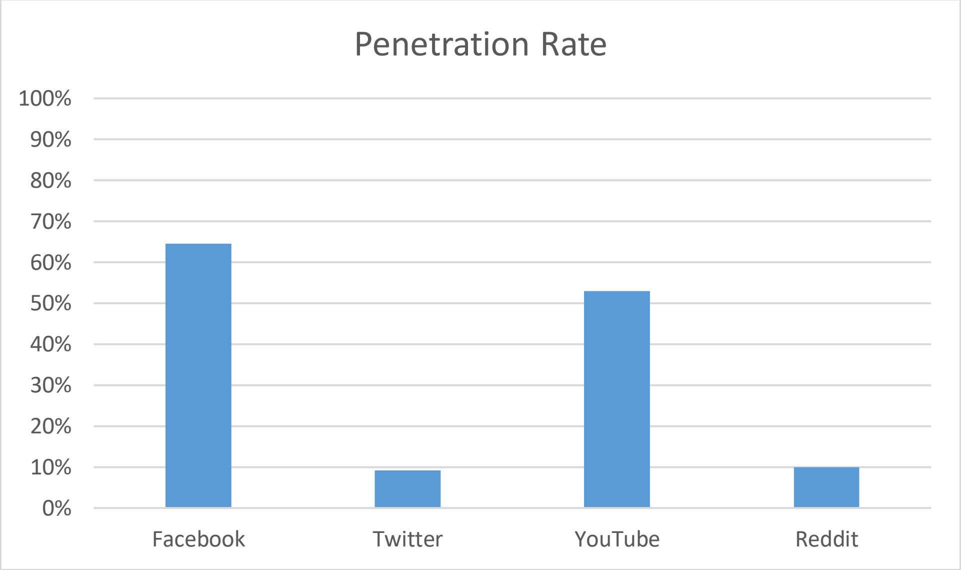


Figure 17. Social Media Penetration Rates

Calculating the initial S.C.A.R.E. scores provides a reference model on the basis that every platform engages the same number of users. Social media penetration rates allow for normalization, which leads to a proper understanding of the likelihood, or propensity, of abuse. The premise is simple: the more people that use something, the greater the risk of misuse. Facebook and YouTube have the largest active user base, therefore, their S.C.A.R.E. score changes less than Reddit's and Twitter's scores. A startling discovery was the worst scoring platform: Facebook. Admittedly Facebook and Twitter tied S.C.A.R.E. scores before normalizing the data. However, after accounting for widespread use, Facebook gained a 10% margin on Twitter. Facebook's high score becomes scary when considering it is poised to dominate the next phase of social networking: the metaverse. Even going so far as to rename the company, Meta. Despite a name change, these flaws exist and will not disappear. Facebook demonstrates the greatest proclivity of exploitation by S.C.A.R.E. actors followed by Twitter, YouTube, and lastly Reddit. Outside of anonymity, Facebook scored the highest or second highest in every metric. Proving that the attention Facebook attracts as the poster child of platform abuse is well-deserved.

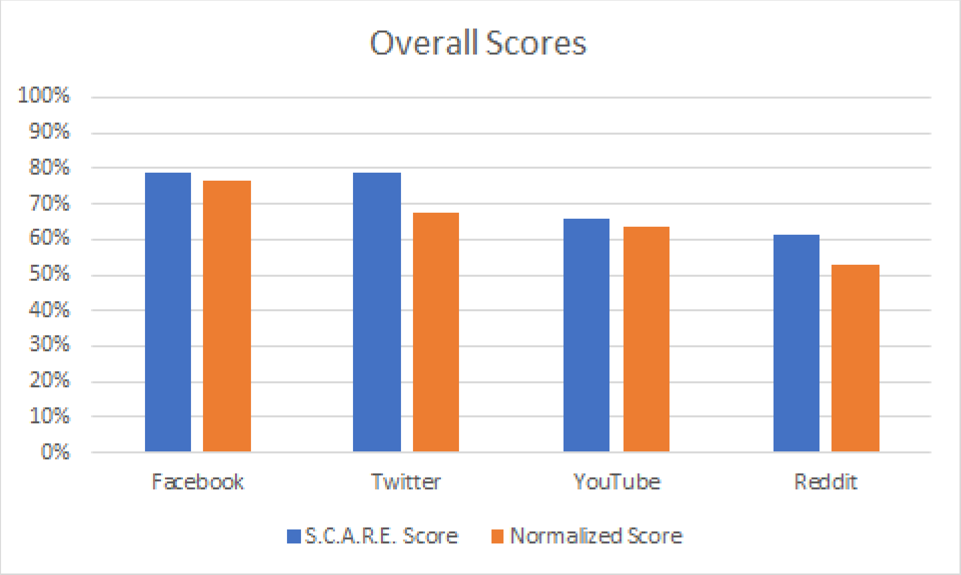
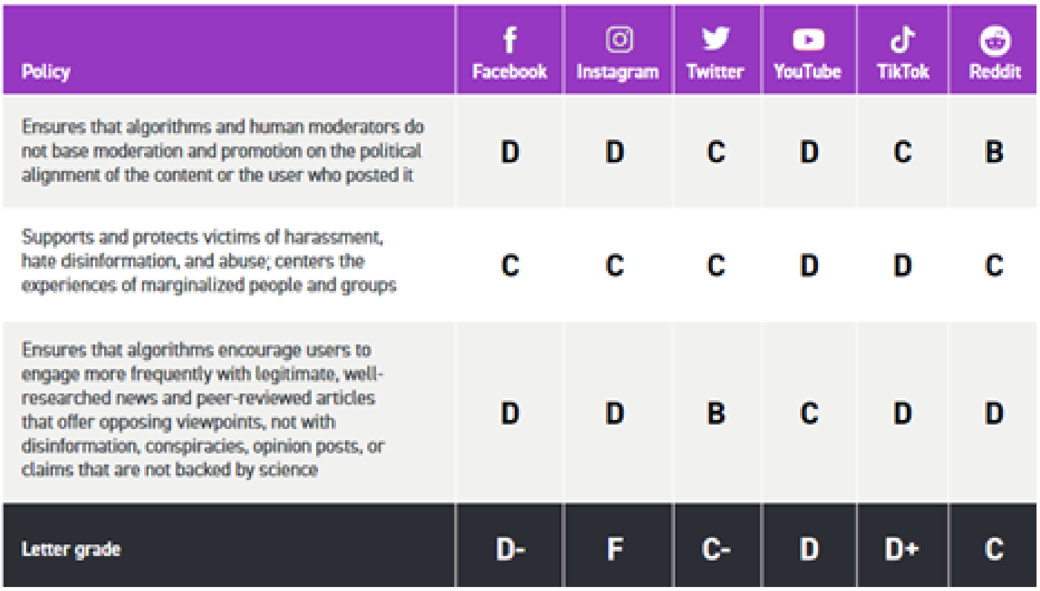


Figure 18. Overall S.C.A.R.E. Scores

These scores represent a likelihood of S.C.A.R.E. abuse. In order to properly calculate risk, the assessor combines these likelihoods with their potential consequences. This aspect falls out of the scope of this thesis and will be addressed in future studies. The likelihood scores remain the focal point of analysis. The S.C.A.R.E. actors grouped together detracts from specific structural features that help scammers more than radicalizers or chaos causers less than election interferers. The higher weighting of personalization and engagement priority served as descriptive of the entirety of S.C.A.R.E. actors but needs alterations for individual evaluations. While the S.C.A.R.E. actors exploit social media platforms through similar weaknesses, the weight placed on each weakness per actor varies. For instance, scammers care most about anonymity, whereas radicalizers care most about personalization. In future studies, these metrics may lead to specific scores for each S.C.A.R.E. actor. This study remains limited to the S.C.A.R.E. collective as a whole abusing social media. If evaluated successfully against a previous study, these scores gain added legitimacy.



Policy	Facebook	Instagram	Twitter	YouTube	TikTok	Reddit
Ensures that algorithms and human moderators do not base moderation and promotion on the political alignment of the content or the user who posted it	D	D	C	D	C	B
Supports and protects victims of harassment, hate disinformation, and abuse; centers the experiences of marginalized people and groups	C	C	C	D	D	C
Ensures that algorithms encourage users to engage more frequently with legitimate, well-researched news and peer-reviewed articles that offer opposing viewpoints, not with disinformation, conspiracies, opinion posts, or claims that are not backed by science	D	D	B	C	D	D
Letter grade	D-	F	C-	D	D+	C

Figure 19. ISD Report

An independent report conducted by the Institute for Strategic Dialogue(ISD) and Ultraviolet assesses the validity of the results. This study ranked the policies and the success of Facebook, Instagram, Twitter, YouTube, TikTok, and Reddit in the areas of content moderation, hate speech, misinformation, etc[11]. Then, the researchers conducted an analysis of policies and instances of violation to assign a letter grade to each platform. Figure 19 displays the last three evaluations and summative scores. The report corroborates the propensity scores assessed for Facebook and Reddit. The ratings of YouTube and Twitter conflicted with the findings of this study. Although they are next to each other in both studies, the ISD report states a substantial difference with YouTube two letter grades lower than Twitter. The content moderation metric could have contributed to this conflict. YouTube's moderation presence was potentially miscalculated as explained in the future studies section. On the other side, Twitter scored artificially high in combating misinformation and stopping hate speech, which starkly contradicts the available literature and official releases by the company. The ISD report awards twitter a ‘C’ on content moderation, while the study found it to be the worst of all the platforms. However, Reddit’s moderation grade agrees with the results in this thesis. For the majority of ratings, the two studies aligned, but the content moderation and overall rankings of Twitter and YouTube bring uncertainty. As a result, the study can not conclusively assess the viability of evaluation metrics. The repetitive conflicts point to the content moderation calculations, which may need modification going forward. Unfortunately, this study found no other social media grading reports or research. It is this author’s hope that new analyses are published to evaluate this topic deeper. It is essential and of timebound importance with the next wave of social networking imminent. If we fail to measure the issues in order to create solutions, the future will be hateful, cruel, authoritarian, and manipulative.

Chapter 7

Future Studies

Areas of Concern

Many major and minor issues arose while conducting this research. These problems spanned the calculation of the B constant, finding valid data, comparing the study's results, and the anonymity metric. The B constant used in the content moderation metric is the first area of concern. The constant represents the standard percentage of all content that should be moderated. The study derived it from combining Facebook Transparency reports and internal reports, which include the specific number of posts moderated and an estimate of how much content the platform misses. Assuming this number stays constant is a major assumption of the thesis. Problems with the content moderation metric did not end there. It proved difficult to find comparative figures for the number of posts made on a platform and the amount moderated. For example, YouTube demonstrated this issue. No overall figures for YouTube moderated versus unmoderated posts, including comments, exists, therefore the study utilized YouTube video statistics only. If the B constant remains constant across all content, this is not an issue. Yet, the expectation that comments yield a greater rate of hate speech than videos introduces a possibility of undercounting.

This lack of reliable data existing about social media platforms affected the accuracy of the results. For example, the statistics concerning the number of posts are not provided by platforms and are instead found by independent firms. Platform transparency reports contain some data points, however, the platform presents this data with horrendous usability. The exact opposite use of captology is found in their networking applications. Navigating the reports proves confusing and disorienting. No data is presented clearly, but rather through a filtered lens that paints the platform in a good light. A definite challenge to correctly computing the content moderation metric. This issue concerning the lack of data transparency by social media companies extends to all metrics. Conflicting statistics and confusing calculations were commonplace across reports. This experience confirms a report completed by the Anti-Defamation League that explains the problems with transparency reports. Critiques fall into one of four categories: hard-to-find reports, inconsistent data across reports, misleading key metrics, and missing necessary context[44]. Essentially, platforms artificially inflate their success at content moderation and fighting abuse through confusing and incoherent reports. Consequently, results are a product of distortion and inconclusive statistics. These variables need better sources moving forward, which hopefully, begin to come from the social media companies themselves with added public pressure.

The third concern outlines the difficulty in comparing this study's findings to others. The most suitable match for the desired purpose was the Institute for Strategic Dialogue's(ISD) utilized in the discussion and analysis section. Both reports score major social media platforms on content moderation efforts, abuse tolerance, and a few other areas that overlap. However, as stated the ISD report relies on opinion-based judgments without numerical backing. Therefore, when the two scores conflicted, neither could be established as more accurate. Despite the two reports agreeing on platform rankings in the majority of cases, no determination on success or failure is made. The next steps aim at finding more substantive ratings to establish whether the metrics in this study prove effective or not. Additionally, increased focus on metric development dwarfs all other priorities. Current metrics describe the weakness they seek to measure but do not fully account for all independent factors. In the case of anonymity, the ability to be anonymous arises from more than the presence of six signup qualities. These factors include IP address tracing, device identifier logs, and social networks, ie friends or connections, on the platform, among others. A key element of future studies centers on the development of comprehensive metrics for anonymity and the three other structural weaknesses.

Improvements to Flaws

For every hole that S.C.A.R.E. actors exploit, there exists a patch that stops them. The solutions require an entire thesis in and of themselves. In fact, this was my original research topic before narrowing it to the first component of finding and measuring S.C.A.R.E. abuse likelihood. Therefore, an introductory level of solutions provides the basis for further research. A combination of successful examples and new ideas starts the dialogue on solving these structural weaknesses.

Anonymity

Anonymity gives S.C.A.R.E. actors the ability to hide in plain sight and conduct their exploitation of platforms and users in complete safety. Stringent and intensive account creation processes substantially decrease these levels of anonymity. A great example is Coinbase's system of verification, which requires a photo ID[51]. Coinbase contends it is necessary to "verify your identity to ensure no one but you changes your payment information" and "prevent fraud"[51]. Understandably, Coinbase is legally required to collect this information as a cryptocurrency broker. However, that should not stop social networking sites committed to battling the S.C.A.R.E. collective as they all claim to. This extra step would reduce anonymous actors and trolls while heightening trust and community online. It is the emblematic emergence of authenticity, the opposite of anonymity, being prioritized. Experts predict the process of ID verification to be increasingly accessible. Especially, as a trend recently emerged of states switching to digital identification capabilities. At the very least this solution provides the substantial hurdle of counterfeiting photo IDs, thus slowing down malicious actors.

Content Moderation

Any choice regarding banning or allowing controversial posts invites criticism. The effects seen in poorly constructed moderation policies showcase contradictions that negate their own points. Restructuring social media as levels circumvents these issues. Instead of one massive community of thought, three levels form a user's online experience. The first resembles the current online environment and errs on the side of not moderating at all. The concept behind this level is sacrifice. It gives users the expectation of scammers, radicalizers, trolls, advertisers, and election interferers in this subsection of the application. The users are given a choice if they want to interact with these malicious actors of their own volition. The second level requires a post to be accepted by both automated content moderation systems and the users of the platform. The self-moderation technique is borrowed from Reddit here. An additional implementation of this concept is on Wikipedia[53]. The model that theoretically should fail but proves its remarkable success in practice. A place where random internet users determine and record facts and history with striking accuracy. Wikipedia triumphs as a self-moderated ecosystem of information. The third level is evaluated for factual truth and assessed by an independent watchdog organization. For this method to work, trust and validity of information must increase as the levels do as well as the strict guidelines of allowed versus not allowed content.

Engagement Priority

The immediate solution for engagement priority, weaved through social media platforms, is unlatching companies' dependence on advertisers' money. Profit generated from other methods exists. For example, the platform could charge users a yearly fee or give them the option to purchase a premium account. The intertwined nature of advertisers and platforms is undeniable. The same is said for users wanting free platforms. A unique opportunity emerges that satisfies both parties: social media users and platforms. Users get rights over their own data. Platforms give every user a choice in regard to how their data is used. If the user chooses to sell their social media data to advertisers, they receive 50% of the revenue and the platform receives the other half. If the user chooses to keep their data private, the platform's limits to obtaining data, selling, data, targeting advertisements, and curating content, are strictly enforced. The overflow effect helps to combat the personalization feature/weakness. Overhauling existing platforms would be a futile fight; therefore, these standards and guidelines shape future systems.

Personalization

As highlighted previously, personalization is traditionally influenced by the engagement priority of a platform. If the advertisement revenue dilemma is properly handled, then targeted advertisements and personalized content should decrease. However, the way a platform protects and handles a user's data matters in every scenario even the one listed above. Additionally, advertisements are less effective when less data is gathered about the user. It becomes a matter of enshrining the right to one's own data. The General Data Protection Regulation in the EU seeks to "protect fundamental rights and freedoms of natural persons and in particular their right to the protection of personal data"[9]. Established in 2018, it has the potential to expand in order to decrease obvious collusion between advertisers and social media. GDPR was a primary driving force behind platforms releasing transparency reports, albeit far from transparent. It needs hefty revisions and improvements, but it signals a change of direction for the better protection of the individual.

Chapter 8

Conclusion

This thesis identifies the underlying mechanisms and structural design flaws that enable a collective of malicious actors, S.C.A.R.E., in exploiting social media users and platforms. The five threat actors: scammers, chaos causers, advertisers, radicalizers, and election meddlers; possess four characteristics: win-lose intention, manipulation of users in a negative way, increased presence with social media, and exploitation of social media structure. In the case studies section, all S.C.A.R.E. actors were found to exploit social media through similar structural weaknesses. The four weaknesses are anonymity, content moderation, engagement priority, and personalization. Then, sufficient evaluation metrics were established to determine a platform's likelihood of abuse by a S.C.A.R.E. actor based on the presence of the four structural issues. Twitter had the greatest propensity for exploitation, however, when accounting for widespread use Facebook displayed a considerably higher likelihood. Due to the lack of available data and prior analyses, the success of this framework and grading social media platforms is inconclusive. Despite this it succeeds in increasing user awareness of captology, S.C.A.R.E., and structural weaknesses found on social media platforms, so that users may reclaim decision-making autonomy. Future research aims at improved metrics and effective solutions that demand better from social networking platforms. Hopefully, attention and accountability force improved reporting and ethics from social media companies. The internet and social networking evolves rapidly. We need to change with it or the consequences will be devastating to civil discourse, society, and a harmonious online ecosystem.

BIBLIOGRAPHY

[1] K. Weill, *Off the edge : flat Earthers, conspiracy culture, and why people will believe anything*. Chapel Hill, North Carolina: Algonquin Books Of Chapel Hill, 2022, p. 159.

[2] Jensen, Michael. *The Use of Social Media by United States Extremists*. July 2018. [Online]. Available: https://www.start.umd.edu/pubs/START_PIRUS_UseOfSocialMediaByUSExtremists_ResearchBrief_July2018.pdf

[3] "Demographics of Social Media Users and Adoption in the United States." *Pew Research Center,* Feb. 2021, www.pewresearch.org/internet/fact-sheet/social-media/#find-out-more.

[4] H. Ritchie and M. Roser, "Technology Adoption," *Our World in Data*, 2017. https://ourworldindata.org/technology-adoption

[5] Bernhard Debatin, Jennette P. Lovejoy, Ann-Kathrin Horn, M.A., Brittany N. Hughes, "Facebook and Online Privacy: Attitudes, Behaviors, and Unintended Consequences," *Journal of Computer-Mediated Communication*, Volume 15, Issue 1, 1 October 2009, Pages 83–108, https://doi.org/10.1111/j.1083-6101.2009.01494.x

[6] D. Milmo and D. Pegg, "Facebook admits site appears hardwired for misinformation, memo reveals," *The Guardian*, Oct. 25, 2021. https://www.theguardian.com/technology/2021/oct/25/facebook-admits-site-appears-hardwired-misinformation-memo-reveals

[7] P. Dizikes, "Study: On Twitter, false news travels faster than true stories," *MIT News | Massachusetts Institute of Technology*, Mar. 08, 2018. https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308

[8] E. Buckels, P. Trapnell, and D. Paulhus, "Trolls just want to have fun," *Personality and Individual Differences*, vol. 67, pp. 97–102, Feb. 2014, doi: 10.1016/j.paid.2014.01.016.

[9] "General Data Protection Regulation (GDPR)," *Intersoft Consulting*, 2013. https://gdpr-info.eu/art-1-gdpr/

[10] R. McNamee, "How to Fix Facebook Before It Fixes Us," *The Washington Monthly,* vol. 50, no. 1–3, pp. 33–40, 2018.

[11] Ultraviolet and Institute for Strategic Dialogue(ISD), "Social Media Report Card," Nov. 2021. [Online]. Available: https://weareultraviolet.org/wp-content/uploads/2021/11/Social-media-report-card.pdf

[12] B. Levick, "Propaganda and the Imperial Coinage," *Antichthon*, vol. 16, pp. 104–116, 1982, doi: 10.1017/s0066477400002999.

[13]Carroll, Abigail. Three Squares: The Invention of the American Meal. United States, Basic Books, 2013

[14] E. Bernays, *Propaganda*. New York, Liveright, 1933. [Online]. Available: https://archive.org/details/BernaysPropaganda/mode/2up

[15] B. J. Fogg, "Persuasive Technology Using Computers to Change What We Think and Do," *Zlibcdn.com*, 2003. https://bunker4.zlibcdn.com/dtoken/fb80cd20d17ef1a786225c78d58fe37f

[16] S. Spies, "Election Interference," *Social Science Research Council*, Apr. 2020, [Online]. Available: https://mediawell.ssrc.org/literature-reviews/election-interference/versions/1-1/

[17] R. Shapiro, "What Your Data Is Really Worth to Facebook," *Washington Monthly*, Jul. 13, 2019. https://washingtonmonthly.com/2019/07/12/what-your-data-is-really-worth-to-facebook/

[18] J. Constine, "Facebook changes mission statement to 'bring the world closer together,'" *TechCrunch*, Jun. 22, 2017. https://techcrunch.com/2017/06/22/bring-the-world-closer-together/

[19] Y. Samrai, "How Stanford Profits Off Addiction," *The Stanford Review*, Feb. 04, 2020. https://stanfordreview.org/how-stanford-profits-tech-addiction-social-media/

[20] R. Kohavi and R. Longbotham, "Online Controlled Experiments and A/B Testing," *Encyclopedia of Machine Learning and Data Mining*, pp. 922–929, 2017, doi: 10.1007/978-1-4899-7687-1_891.

[21] S. Lambert, "Number of Social Media Users in 2020: Demographics & Predictions - Financesonline.com," *financesonline.com*, 2020. https://financesonline.com/number-of-social-media-users/

[22] D. Kemps and E. Ekins, "Poll: 75% Don't Trust Social Media to Make Fair Content Moderation Decisions, 60% Want More Control over Posts They See," *Cato.org*, Dec. 15, 2021. https://www.cato.org/survey-reports/poll-75-dont-trust-social-media-make-fair-content-moderation-decisions-60-want-more#american-support-tech-industry

[23]"Consumer Protection Data Spotlight | Scams starting on social media proliferate in early 2020," Oct. 2020. [Online]. Available: https://www.ftc.gov/system/files/attachments/blog_posts/Scams%20starting%20on%20social%20media%20proliferate%20in%20early%202020%20/data_spotlight_oct_2020.pdf

[24]"SCAMMER | meaning in the Cambridge English Dictionary," *Cambridge.org*, 2019. https://dictionary.cambridge.org/dictionary/english/scammer

[25] Y. Talib and F. Rusly, "Falling Prey for Social Media Shopping Frauds: The Victims' Perspective," Oct. 2015. [Online]. Available: https://repo.uum.edu.my/id/eprint/17543/1/27_ICoEC2015%20185-189.pdf

[26] SEC, “Internet and Social Media Fraud | Investor.gov,” *www.investor.gov*. https://www.investor.gov/protect-your-investments/fraud/types-fraud/internet-and-social-media-fraud

[27] *Violence And Trolling On Social Media.* Amsterdam: Amsterdam University Press, 2020. [Online]. Available: https://library.oapen.org/viewer/web/viewer.html?file=/bitstream/handle/20.500.12657/42883/9789048542048.pdf

[28] Statista, “Online or social media targeting effectiveness 2019,” *Statista*, Sep. 24, 2021. https://www.statista.com/statistics/303726/social-media-targeting-effectiveness/

[29] E. Terkki, A. Rao and S. Tarkoma, "Spying on Android users through targeted ads," *2017 9th International Conference on Communication Systems and Networks (COMSNETS)*, 2017, pp. 87-94, doi: 10.1109/COMSNETS.2017.7945362.

[30] S. Goodson, “If You’re Not Paying For It, You Become The Product,” *Forbes*, Mar. 05, 2012. https://www.forbes.com/sites/marketshare/2012/03/05/if-youre-not-paying-for-it-you-become-the-product/?sh=66214e55d6ee

[31] R. Borum, “Rethinking Radicalization,” *Journal of Strategic Security*, vol. 4, no. 4, pp. 1–6, 2011, [Online]. Available: https://www.jstor.org/stable/26463909?seq=2

[32] E. Brooks, “Jan. 6 panel subpoenas ‘misinformation’ and ‘extremism’ records from YouTube, Facebook, Reddit, and Twitter,” *Washington Examiner*, Jan. 13, 2022. https://www.washingtonexaminer.com/news/january-6-select-committee-subpoena-youtube-facebook-reddit-twitter

[33] Jens David Ohlin, *Election interference : international law and the future of democracy*. Cambridge, United Kingdom ; New York, Ny: Cambridge University Press, 2020. [Online]. Available: https://www.cambridge.org/core/services/aop-cambridge-core/content/view/72535E08C80FBBE6601BEEC666F7CDBC/9781108494656c1_10-39.pdf/what_is_election_interference.pdf

[34] J. Kosseff, “A User’s Guide to Section 230, and a Legislator’s Guide to Amending It (or not).” Aug. 2021. [Online]. Available: https://deliverypdf.ssrn.com/delivery.php?ID=544017099074094123031119097017069011027055029016031058025074023064120016095093111074110122119106047059058117072126126122080030102009075041077098111025125111071029006031016046086008124117007122115117122102030100099121019126098088091029096113004101084105&EXT=pdf&INDEX=TRUE

[35] 47 U.S.C. § 230(c)(1).

[36] 47 U.S.C. § 230(c)(2)(A)(B).

[37] House of Representatives, Subcommittee on National Security, Committee on Oversight and Government Reform, “Radicalization: Social Media and the Rise of Terrorism.” Oct. 28, 2015. [Online]. Available: https://docs.house.gov/meetings/GO/GO06/20151028/104134/HHRG-114-GO06-Transcript-20151028.PDF

[38] J. Isaak and M. J. Hanna, "User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection," in *Computer*, vol. 51, no. 8, pp. 56-59, August 2018, doi: 10.1109/MC.2018.3191268. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8436400

[39] *Social Media (Anti-Trolling) Bill 2022*. 2022. [Online]. Available: https://www.legislation.gov.au/Details/C2022B00015

[40] Meta, “Community Standards Enforcement Report,” 2021. [Online]. Available: https://transparency.fb.com/data/community-standards-enforcement/

[41] Facebook, “Facebook,” *Facebook*, 2021. https://www.facebook.com/

[42] C. Smith, “How many Facebook posts are made each minute?” https://blog.dlvrit.com/how-many-facebook-posts-are-made-each-minute-infographic/

[43]“https://twitter.com/i/flow/signup,” *Twitter*. https://twitter.com/i/flow/signup

[44] K. Sorenson, “What’s Wrong with Transparency Reporting (and How to Fix It),” *Anti-Defamation League*, 2022. https://www.adl.org/resources/reports/whats-wrong-with-transparency-reporting-and-how-to-fix-it

[45] D. Albright, “Facebook Ads CTR: Benchmarks and Best Practices,” *Databox*, Jul. 16, 2019. https://databox.com/average-facebook-ctr

[46] N. Reiff, “How Does Twitter Make Money?,” *Investopedia*, 2022. https://www.investopedia.com/ask/answers/120114/how-does-twitter-twtr-make-money.asp

[47] B. Sharma, “575K Tweets, 5.7 Mn Google Searches & More, Here’s What Happens On The Internet Every Minute,” *IndiaTimes*, Nov. 27, 2021. https://www.indiatimes.com/technology/news/what-happens-on-the-internet-every-minute-575k-tweets-57-million-google-searches-more-555328.html

[48] J. Barker, “Social Advertising Benchmarks for 2022,” *Brafton*, Dec. 21, 2021. https://www.brafton.com/blog/social-media/social-advertising-benchmarks/

[49] T. Rose, “How YouTube Makes Money ($19.8 Billion in Revenue) | Business Model,” *Entrepreneur360*, Aug. 30, 2021. https://entrepreneur-360.com/how-does-youtube-make-money-21501

[50] Dennis, “YouTube Ads Benchmarks (2021),” *Store Growers*, Jan. 18, 2021. https://www.storegrowers.com/youtube-ads-benchmarks/

[51] “Identity document verification | Coinbase Help,” *help.coinbase.com*. https://help.coinbase.com/en/coinbase/getting-started/getting-started-with-coinbase/id-doc-verification

[52] M. Osman, “Wild and Interesting Facebook Statistics and Facts (2019),” *Kinsta*, Dec. 28, 2018. https://kinsta.com/blog/facebook-statistics/

[53] O. Benjakob and R. Aviram, “A Clockwork Wikipedia: From a Broad Perspective to a Case Study,” *Journal of Biological Rhythms*, vol. 33, no. 3, pp. 233–244, Apr. 2018, doi: 10.1177/0748730418768120.

[54] Statista, “Facebook ad revenue 2009-2018,” *Statista*, Feb. 05, 2021. https://www.statista.com/statistics/271258/facebooks-advertising-revenue-worldwide/

[55] Statista, “Facebook: annual revenue 2018 | Statistic,” *Statista*, 2018. https://www.statista.com/statistics/268604/annual-revenue-of-facebook/

[56] A. Badawy, E. Ferrara and K. Lerman, "Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign," *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2018, pp. 258-265, doi: 10.1109/ASONAM.2018.8508646.

[57] Statista, “Global Social Media Ranking 2021,” *Statista*, Mar. 08, 2022. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/
[58] “reddit.com: Join the worldwide conversation,” *Reddit.com*, 2022. https://www.reddit.com/register/

[59] eMarketer Editors, “Reddit Is on Pace to More Than Double Its Ad Revenues By 2021,” *eMarketer*, Mar. 26, 2019. https://www.emarketer.com/content/reddit-to-cross-100-million-in-ad-revenues-in-2019

[60] D. Curry, “Reddit Revenue and Usage Statistics (2021),” *Business of Apps*, Jan. 11, 2022. https://www.businessofapps.com/data/reddit-statistics/

[61] S. Zannettou, T. Caulfield, W. Setzer, M. Sirivianos, G. Stringhini, and J. Blackburn, “Who Let The Trolls Out? Towards Understanding State-Sponsored Trolls,” Nov. 2018. [Online]. Available: https://arxiv.org/pdf/1811.03130v1.pdf

[62] “insurance at DuckDuckGo,” *duckduckgo.com*. https://duckduckgo.com/?t=ffab&q=insurance&atb=v302-1&ia=web

[63] R. Warren, *Targeted and Trolled : the Reality of Being a Woman Online (an Original Digital Short)*. London: Transworld Digital, 2015.

[64] L. Rainie, “Americans’ complicated feelings about social media in an era of privacy concerns,” *Pew Research Center*, Mar. 27, 2018. https://www.pewresearch.org/fact-

tank/2018/03/27/americans-complicated-feelings-about-social-media-in-an-era-of-privacy-concerns/

[65] L. Santhanam, “American voters worry they can’t spot misleading information, poll finds,” *PBS NewsHour*, Jan. 21, 2020. https://www.pbs.org/newshour/politics/social-media-disinformation-leads-election-security-concerns-poll-finds

[66] B. Green, “What Is the Main Purpose of Social Media: Entertainment or Education?,” *Markkula Center for Applied Ethics*, Dec. 16, 2019. https://www.scu.edu/ethics-spotlight/social-media-and-democracy/what-is-the-main-purpose-of-social-media-entertainment-or-education/

[67] M. Roser and M. Nagdy, “Nuclear Weapons,” *Our World in Data*, 2013. https://ourworldindata.org/nuclear-weapons

[68] [C. Brucker, “The Rise of Marketing and Advertising,” *Business Enterprise in American History*. Houghton Mifflin, The University of Michigan, 1986. [Online]. Available: https://faculty.atu.edu/cbrucker/Engl5383/Marketing.htm

[69] L. Harding, “How Bernays Changed the World Through PR,” *Frontier Centre for Public Policy*, Nov. 09, 2021. [Online]. Available: https://fcpp.org/2021/11/09/how-bernays-changed-the-world-through-pr/

[70] J. Balkin, “How to Regulate (and Not Regulate) Social Media,” presented at the Association for Computing Machinery Symposium on Computer Science and Law, New York City, Mar. 2020. [Online]. Available: https://knightcolumbia.org/content/how-to-regulate-and-not-regulate-social-media

[71] T. Johnson, “The FCC’s Authority to Interpret Section 230 of the Communications Act,” *Federal Communications Commission*, Oct. 21, 2020. https://www.fcc.gov/news-events/blog/2020/10/21/fccs-authority-interpret-section-230-communications-act

[72] J. Patchin, “Summary of Our Cyberbullying Research (2004-2016),” *Cyberbullying Research Center*, Jul. 10, 2019. https://cyberbullying.org/summary-of-our-cyberbullying-research

[73] Google, “Google Trends - ‘election interference,’” *Google Trends*. https://trends.google.com/trends/explore?date=all&geo=US&q=election%20interference

[74] T. Wood, “Visualizing the Evolution of Global Advertising Spend (1980-2020),” *Visual Capitalist*, Nov. 10, 2020. https://www.visualcapitalist.com/evolution-global-advertising-spend-1980-2020/

[75] S. Martín, “Facebook engagement: What it is and ways to increase it,” *Metricool*, Jun. 20, 2019. https://metricool.com/what-is-facebook-engagement/

[76] Y. Y. A. Talib, “The Current State of Social Commerce Fraud in Malaysia and the Mitigation Strategies,” *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 2, pp. 1593–1599, Apr. 2020, doi: 10.30534/ijatcse/2020/105922020.

[77] S. Ahmad, “‘It’s Just the Job’: Investigating the Influence of Culture in India’s Commercial Content Moderation Industry,” MSc Thesis, University of Oxford, 2018. [Online]. Available: https://osf.io/preprints/socarxiv/hjcv2/

[78] C. V. Baccarella, T. F. Wagner, J. H. Kietzmann, and I. P. McCarthy, “Social media? It’s serious! Understanding the dark side of social media,” *European Management Journal*, vol. 36, no. 4, pp. 431–438, Aug. 2018, doi: 10.1016/j.emj.2018.07.002.

[79] “Transparency Report 2019 - Reddit,” *Reddit.com*, 2019. https://www.redditinc.com/policies/transparency-report-2019

[80] F. Allen, “Mum’s heartbreaking letter to bullies who drove her son to suicide after he was teased for not playing Call of Duty,” *The Sun*, Oct. 05, 2016. https://www.thesun.co.uk/news/1916122/mum-pens-heartbreaking-letter-to-bullies-who-drove-her-son-to-suicide-after-he-was-teased-for-not-being-allowed-to-play-call-of-duty/

[81] S. Balloo, “Mum’s message to parents after cyberbullied son took his own life,” *Manchester Evening News*, Oct. 12, 2017. https://www.manchestereveningnews.co.uk/news/uk-news/cyberbullying-social-media-warning-parents-13751702

[82] H. King and S. A. O’Brien, “Twitter CEO Dick Costolo quits,” *CNNMoney*, Jun. 11, 2015. https://money.cnn.com/2015/06/11/technology/twitter-ceo-dick-costolo-quits/index.html

[83] A. Hern, “Twitter CEO: We suck at dealing with trolls and abuse,” *The Guardian*, Feb. 05, 2015. https://www.theguardian.com/technology/2015/feb/05/twitter-ceo-we-suck-dealing-with-trolls-abuse

[84] Exploring Online Anonymity, “Case study 3: The detrimental effects of trolling online,” *Medium*, Oct. 29, 2019. https://medium.com/@shannonwilkins_93340/case-study-3-the-detrimental-effects-of-trolling-online-c0b67e89204d

[85] A. Deb, S. Donohue, and T. Glaisyer, “Is Social Media a Threat to Democracy?,” Oct. 2017. [Online]. Available: https://www.omidyargroup.com/wp-content/uploads/sites/7/2017/10/Social-Media-and-Democracy-October-5-2017.pdf

[86] C. Zakrzewski, “microtarget,” *Merriam Webster*. [Online]. Available: https://www.merriam-webster.com/dictionary/microtarget

[87] A. Gotter, “What in the World … ! Surprising Facts About Reddit Ads,” *Agorapulse*, Sep. 09, 2021. https://www.agorapulse.com/blog/facts-reddit-ads/

[88] S. Kemp, “Digital 2021 April Statshot Report,” Kepios, Apr. 2021. [Online]. Available: https://datareportal.com/reports/digital-2021-april-global-statshot

[89] “What Is a ‘Good’ Click-Through Rate? | CTR Benchmarks,” *CXL*, 2022. https://cxl.com/guides/click-through-rate/benchmarks/

[90] Google, “YouTube Community Guidelines enforcement,” 2021. [Online]. Available: https://transparencyreport.google.com/youtube-policy/appeals?hl=en&total_removed_videos=period:2021Q4;exclude_automated:all&lu=total_removed_videos&total_videos_reinstated=period:2020Q4

[91] L. Ceci, “Youtube average video length by category 2018,” *Statista*, Aug. 23, 2021. https://www.statista.com/statistics/1026923/youtube-video-category-average-length/

[92] R. Jindal, A. Falah, A. Anwar, and M. Ahmed, “Privacy-Preserving Analytics for Social Network Data,” in *Securing Social Networks in Cyberspace*, A.-S. K. Pathan, Ed. CRC Press, 2021, pp. 17–33.

[93] B. Beyersdorf, “1061 Regulating the ‘Most Accessible Marketplace of Ideas in History’: Disclosure Requirements in Online Political Advertisements After the 2016 Election,” *California Law Review*, vol. 107, no. 3, Jun. 2019, [Online]. Available: https://www.californialawreview.org/print/regulating-the-most-accessible-marketplace-of-ideas-in-history-disclosure-requirements-in-online-political-advertisements-after-the-2016-election/

[94] R. Sharp, “People join extremist groups on Facebook due to algorithms,” *Daily Mail*, May 27, 2020. https://www.dailymail.co.uk/news/article-8360073/More-60-people-joining-extremist-groups-Facebook-pages-recommended-algorithms.html

[95] N. HUBI, “The Role of Social Media in Influencing Radicalization,” MSc Thesis, United States International University-Africa, 2019. [Online]. Available: http://erepo.usiu.ac.ke/bitstream/handle/11732/4915/NESTEHA%20HUSSEIN%20MOHAMED%20HUBI%20MAIR%202019.pdf

[96] E. Buçaj, “The role of internet and the new dimension of computer terrorism,” University “Ukshin Hoti” Prizren, 2021. [Online]. Available: https://www.researchgate.net/publication/348469361_The_role_of_internet_and_the_new_dimension_of_computer_terrorismerrorism

[97] J. D’Onfro, “Google’s new rules for employees: No doxxing, trolling or name calling,” *CNBC*, Jun. 27, 2018. https://www.cnbc.com/2018/06/27/google-issues-new-policy-on-harassment-discrimination-retaliation.html

[98] B. Gordon and W. Hartmann, “Advertising Effects in Presidential Elections,” *Marketing Science*, vol. 32, no. 1, pp. 19–35, Nov. 2012, [Online]. Available: https://doi.org/10.1287/mksc.1120.0745

[99] Google, “Create your Google Account,” *Google.com*. https://accounts.google.com/signup/v2/webcreateaccount

[100] “Twitter Transparency Center,” *Twitter Transparency Center,* 2021. https://transparency.twitter.com

[101] S. Aslam, “YouTube by the Numbers (2019): Stats, Demographics & Fun Facts,” *Omnicoreagency.com*, Sep. 05, 2019. https://www.omnicoreagency.com/youtube-statistics/

[102] “MANIPULATION | Definition in the Cambridge English Dictionary,” *dictionary.cambridge.org*, dictionary.cambridge.org/us/dictionary/english/manipulation.

[103] M. Cartwright, “Roman Coinage,” *World History Encyclopedia*, Apr. 19, 2018. https://www.worldhistory.org/Roman_Coinage/

[104] Center for Tobacco Products, “Overview of the Family Smoking Prevention and Tobacco Control Act,” *U.S. Food and Drug Administration*, 2019. https://www.fda.gov/tobacco-products/rules-regulations-and-guidance/family-smoking-prevention-and-tobacco-control-act-overview

[105] B. Baruch, T. Ling, R. Warnes, and J. Hofman, “Violent Extremism Evaluation Measurement (VEEM) Framework,” *www.rand.org*, Oct. 2018. https://www.rand.org/randeurope/research/projects/violent-extremism-evaluation-measurement-framework-veem.html

[106] C. Baccarella, T. Wagner, J. Kietzmann, and I. McCarthy, “Social media? It’s serious! Understanding the dark side of social media,” *European Managment Journal*, vol. 36, pp. 431–438, 2018, doi: 10.1016/j.emj.2018.07.002.

[107] CEP Staff, “Germany’s NetzDG Content Moderation Law Undergoes Revamp,” *Counter Extremism Project*, Oct. 07, 2020. https://www.counterextremism.com/blog/germanys-netzdg-content-moderation-law-undergoes-revamp

[108] P. Litras, “The new Social Media (Anti-Trolling) Bill: Will it work?,” *Federation University*, May 13, 2022. https://federation.edu.au/news/articles/the-new-social-media-anti-trolling-bill-will-it-work

[109] Statista, “Global Social Media Ranking 2021,” *Statista*, Mar. 08, 2022. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[110] B. Popik, “‘There is no truth in news and no news in truth’ (Russian Izvestia and Pravda adage),” *www.barrypopik.com*, Aug. 10, 2016.

https://www.barrypopik.com/index.php/new_york_city/entry/there_is_no_truth_in_news_and_no_news_in_truth

[111] C. E. Ackerman, “Big Five Personality Traits: The OCEAN Model Explained [2019 Upd.],” *PositivePsychology.com*, Jun. 19, 2019. https://positivepsychology.com/big-five-personality-theory/

[112] T. Soomro and S. Irshad, “Identity Theft and Social Media,” *International Journal of Computer Science and Network Security*, vol. 18, no. 1, Jan. 2018.

[113] K. Matthews, “Social movements and the (mis)use of research: Extinction Rebellion and the 3.5% rule,” *Interface: a journal for and about social movements*, vol. 12, no. 1, pp. 591–615, Jul. 2020, [Online]. Available: https://www.researchgate.net/profile/Kyle-Matthews/publication/346512007_Social_movements_and_the_misuse_of_research_Extinction_Rebellion_and_the_35_rule/links/5fc5c4dea6fdcce9526914eb/Social-movements-and-the-misuse-of-research-Extinction-Rebellion-and-the-35-rule.pdf

[114] H. Tankovska, “Social networks: penetration in selected countries 2019 | Statistic,” *Statista*, 2019. https://www.statista.com/statistics/282846/regular-social-networking-usage-penetration-worldwide-by-country/

[115] J. Khalili, “How Reddit turned its millions of users into a content moderation army,” *TechRadar*, Jun. 19, 2021. https://www.techradar.com/news/how-reddit-turned-its-millions-of-users-into-a-content-moderation-army

ACADEMIC VITA

ACADEMIC VITA

AUSTIN THOET

Denver, CO 80238 • (720) 690-7141 • austinthoet@gmail.com

THE PENNSYLVANIA STATE UNIVERSITY, University Park, PA
Aug 2022
SCHREYER HONORS COLLEGE **-** Major in Security Risk Analysis; NSA Certificate of Achievement

EC-COUNCIL - Certified Ethical Hacker (CEH) Certification
March 2021

WORK EXPERIENCE

Production Scheduling and Software Development, **EFI POLYMERS,** Denver, CO
May-Aug 2021

- Scheduled product batches based on mixer availability, chemistry, and raw materials, among other factors to meet customer orders in a timely manner
- Modernized and standardized Access database utilized for products internally and externally

Marketing and SEO Analyst Intern, **STIDDLE,** San Francisco
May-Aug 2020

- Complete technical analysis to find powerful terms that may be utilized for customer targeting
- Web scraping with python, BeautifulSoup, etc to acquire and compile data

Online Tutoring Coordinator, **PENN STATE UNIVERSITY,** University Park, PA
Jan-Dec 2019

- Ran the online tutoring program for Penn State's College of Information Science and Technology, which included content creation, presentations, scheduling, interviewing, and online/in-person tutoring

Writer, **NATO SCHOOL,** Oberammergau, Germany
Sep-Dec 2018

- Developed a module on emerging threats in technology for the course curriculum of NATO officers

Mobile Applications Intern, **ACCUWEATHER,** University Park, PA
Aug-Dec 2018

- Utilized C# and various tools to automate tasks
- Fixed bugs and added features to the flagship AccuWeather iOS application

Software Engineering Intern, **LOCKHEED MARTIN,** Rockville, MD
May-Aug 2018

- Overhauled a $100 million domestic/international consultant management application
- Implemented on .Net framework with Visual Basic, SQL, HTML/CSS, Javascript, jQuery
- Gained experience with Agile Methodology, Jira, and the Software Development Life Cycle

Software Engineer, **PSU HOSPITALITY MANAGEMENT,** University Park, PA
Aug 2017-Aug 2018

- Developed a finance-management Android and iOS application
- Implemented Google's Firebase UI, Authentication, and Real-time Database

PERSONAL DEVELOPMENT & OTHER WORK EXPERIENCE

Line Cook, **THE TAVERN,** University Park, PA
Sep 2021-May 2022

Gaucho Chef, **FOGO DE CHAÕ**, Denver, CO
Jan-Nov 2020

President, **NO LOST GENERATION,** University Park, PA
Feb 2018-Dec 2019

- Lead an organization that supports refugees through fundraising events, awareness campaigns, and academic programming

Member, **COMPETITIVE CYBERSECURITY ORGANIZATION,** University Park, PA
Aug 2017-Sep 2019

- Competed in the National Cyber League and CCDC competitions among other challenges

Cadet, **AIR FORCE RESERVE OFFICER TRAINING CORPS,** University Park, PA
Aug 2018-Feb 2019

- Refined numerous disciplines including leadership, briefings, professionalism, and physical fitness

SKILLS & AWARDS

Proficient: Java, Swift, SQL, Python, JavaScript, C++, R
Familiar with: Visual Basic, C#, Ruby, C, Kotlin
Languages: German Language Certificate DSD I and II (B1/B2 level); Basic Spanish
Other Interests: cooking, smoking meat, soccer, reading, bus conversion, hiking, skiing