Redefining Censorship in the Digital Age

by

Kaylee Henry

Presented in partial fulfillment of the requirements

for

Departmental Honors

in the

Department of English and Communication Arts

(Communication Arts)

Hood College

April 2020

Abstract

This paper explores the First Amendment and its role in online communication. Three main social media networks Facebook, Instagram and Twitter are examined throughout this study. This paper examines these social networking site's legal obligations, or lack thereof, to the First Amendment. Supreme court cases as well as smaller, isolated, cases are outlined. Three main categories will be covered as follows: fake news, shadow banning, artificial intelligence and law.

Redefining Censorship in the Digital Age

Digital technology has broken down the walls that separate the business world – which consumes our professional lives – with the everyday activities of our personal lives. There is no longer a switch we can press to stop one life and begin the other. A more startling synthesis, however, may be occurring that has potential for greater impact. A person lives not only a physical life, but, thanks to digital technology, also leads a virtual life. Much like the merging of the workplace and personal lives, are our physical and virtual lives becoming one?

While the scope of this research will focus on social media and the power that private business has on digital censorship, the digital age is transforming much more and innovating everyday objects like refrigerators, washing machines, vacuums and even door locks into smart devices. Smart devices are indirectly designed to increase not only life expectancy but quality of life of the user. It's helped people in all generations. Smart medical equipment can be put in homes which sends all the information to a doctor that may be far away. Simple tasks like grocery shopping to turning on the thermostat can be automated. Life is being lived and controlled through a smart device.

Social media has become an integral aspect of life, nowhere more evident than in American society. It requires user-generated content. Social media sites are used to connect us with family and friends, watch short clips for entertainment and to catch up on the most recent news. It's an instantaneous form of communication that can occur within small communities or a global audience. In fact, 72% of Americans use some type of social media which is a 67% increase from 2005 when social media began to be widely used (PEW, 2019). It's become a daily routine to where life revolves around your social network. Wake up, check Twitter. Make breakfast, post on Facebook. Get to work, reply on Instagram.

There is an age demographic difference in the use of platforms for citizens of the United States. Facebook, which launched in 2004, is a site where users can post links to content on the web, share images, comment and view other user's posts, is primarily used by the 50 to 64 age range at 68% of users. Newer sites like Instagram, which launched in 2010, a photo-sharing social network, and Twitter, which launched in 2006, a site where users post and interact with messages called "tweets", are used in the younger demographic range of ages 18 to 29-year-olds at 67% and 62% (Perrin & Anderson, 2019).

Twitter is the epitome of online microblogging. Twitter users share short messages that are 280 characters or less referred to as "tweets." Tweets can be about any given subject, from jokes to breaking news and can be accessed via either smart device or laptop. The content can include text, images, gifs or videos. These messages are shared with other users who have opted in to follow the sender. Tweets can also be seen by non-followers through the use of hashtags or by following specific topics.

Twitter does not claim responsibility for the content its users choose to share and also, "Reserve[s] the right to remove content that violates the user agreement, including for example, copyright or trademark violations, impersonation, unlawful conduct, or harassment" (Twitter, "Twitter Terms of Service"). By not claiming ownership of its users' content, the user, therefore, owns the content. This also gives Twitter the, "Worldwide, non-exclusive, royalty-free license to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute," the content (Twitter, "Twitter Terms of Service"). At any given time, a user can terminate the legal agreement by deactivating the account.

Facebook is a social networking site where users create free accounts to connect with friends, family and acquaintances through sharing images, music, videos, links and thoughts and

opinions. A Facebook user must send a friend request to another user in order to see what he or she posts, and vice versa. The following must be mutual, unlike Twitter. Aside from personal profiles that require friend requests, Facebook has group pages, fan pages and business pages that allow businesses, organizations and groups to gather to share thoughts and market goods and services without becoming friends.

Each user must agree to the community standards and can be removed if those standards are violated. Content will be taken down if the material is not: “authentic and is misrepresenting who someone is or what they are doing, safe and intimidates, excludes or even silences others, private and violates personal privacy, dignity and harasses or degrades other individuals” (Facebook, "Community Standards").

Like Twitter, Facebook has permission to use the content created and shared on the platform. Upon joining, in the terms and conditions the user signs off to give Facebook, “a non-exclusive, transferable, sub-licensable, royalty-free, and worldwide license to host, use, distribute, modify, run, copy, publicly perform or display, translate, and create derivative works of your content” (Facebook, "Terms of Service"). This means that Facebook can take an image posted, store it and share it with others. This agreement between both parties ends when the content is deleted from Facebook’s systems.

Instagram is one of the Facebook Products; therefore, the terms of use for Instagram is an agreement between the user and Facebook. It is the same terms of use as if the user was signing up for Facebook. Where Facebook can be seen as more text-based, Instagram is driven through the use of images and videos. Despite the emphasis on visual sharing, the basis of the platform is the same as the two mentioned previously. A person creates an account, shares content and

follows other personal and/or business accounts to see their posts. The user can interact with other accounts by not only following but by liking, commenting, messaging and even tagging.

With every seven-in-ten Americans using social media, it may seem like the foolproof way to exercise one's First Amendment rights (PEW, 2019). It is, however, counterintuitive as the two's purposes do not align. So, what is the First Amendment and what does it mean for free speech on the internet in the United States? The First Amendment protects the freedom of speech, religion and the press as well as the right to peaceful protest and to petition the government. The meaning of this has been subject to continuing interpretation since it was adopted in 1791.

The First Amendment's protections are not absolute. There are forms of speech and expression that are not protected that have been established through cases adjudicated by the Supreme Court. Some of the expressions and/or speech that are not protected include but are not limited to child pornography established by the court's decision in New York v. Ferber, libel and slander established by the court's decision in New York Times Co v. Sullivan, the burning of draft cards as an anti-war protest established by the court's decision in United States v. O'Brien, the creation or distribution of obscene materials established by the court's decision in Miller v. California and words or actions meant to incite violence or influence others to commit acts of violence established by the court's decision in Schenck v. United States.

Censorship occurs when individuals or groups try to delete or limit information and ideas that are disseminated throughout society. In other words, the person attempting to distribute this information and ideas is required to have someone examine the content before it's released. The information and/or ideas can be in the form of words, images or ideas and are typically found offensive by the individual or group doing the censorship. Most commonly, what is attempting to

be censored is protected under the First Amendment which is what causes outrage amongst society. The government has the right to censor any form of medium that is not protected under the First Amendment; it is only illegal when things that are protected are attempted to be censored. While there are stipulations for government censorship, private businesses have the right to censor anything without First Amendment implications.

Though private businesses like Facebook, Instagram and Twitter have no legal obligation to follow the First Amendment, a couple of the main issues that social media users find themselves involved in and concerned with are copyright infringement and the definition of obscenity.

To understand what copyright infringement is, one must understand what intellectual property is. Intellectual property, “Refers to creations of the mind, such as inventions; literary and artistic works; designs; and symbols, names and images used in commerce,” (WIPO, "Types of intellectual property"). Copyright is the legal right of the owner of the intellectual property to have ownership and decide what can be done with that content. It protects the owner(s) of the original material from unauthorized duplication or use. Copyright infringement happens when the original material is used without the permission of the owner. The individual taking the content at the expense of the owner’s profit is more likely to be sued for copyright infringement than someone taking the content with no potential for profit.

The model for social media is to share; share videos, share images and share text. With this model, it’s easy for users to fall subject to copyright infringement. Downloading an image and resharing it without the creator’s permission can constitute infringement.

Meantime, what’s obscene to one person may not be seen as obscene to the next. That is where the definition of obscenity becomes quite objective. Obscenity is defined by “offensive to

morality or decency; indecent; depraved," (Dictionary.com, "Obscene"). Most social media platforms will have their own definitions of what is considered obscene. Typically, the platform will identify categories of obscene language, which again, is unintentionally open to interpretation from person to person based on a person's beliefs.

While there are fairly clear-cut limitations to the freedoms outlined in the First Amendment, the internet has created plenty of confusion. The issue isn't that the First Amendment is not enforced online; it is. The issue is that people are using primarily privately-operated websites like Facebook, Twitter and Instagram to communicate as if they were public utilities where free speech is protected. It's a vital component of these companies' business models to make their users feel free to express themselves and encourage more user participation. In reality, these companies have no responsibility to free speech because they are private corporations. That leaves us with, how can free speech be protected when there are private companies with a large mass of users who do not have to obey a set of laws?

Fake News

Prior to the rise of social networking sites, people got their news from trustworthy sources, i.e., news outlets that follow strict codes of conduct. The model of user-generated content for social media sites created a way to disseminate information with very little to no regulation because there are no strict codes of conduct or law to follow. Information that has no validity and/or the majority of the content is false is what defines fake news. Many social media sites are attempting to implement policies to remove this type of content. In essence, they are trying to censor it.

Fake information, while not a new term, thrives on its ability to disguise itself and appear believable to audiences. A lack of understanding the motives or abilities of the users who

generate content contribute to the increase of its spread. Not all fake news is created alike, there are different types made for different purposes. There are six major easily identifiable types which are clickbait, propaganda, satire/parody, sloppy journalism, misleading headlines and biased/slanted news (Web Wise, 2020). Some of these types can be interchangeable and therefore multiple types can be found within one medium.

Clickbait is done deliberately to get the user to click a(n) article, image or video. Most often it's in the form of a sensationalized headline, which may include a photo, that appeals to a person's emotions and curiosity. While clickbait isn't inherently bad, when used in fake news, the content within the sensationalized headline typically doesn't live up to what it appears to be, leaving the user either misinformed or disappointed; meanwhile, the creators are profiting off the clicks of the link. Propaganda is used to promote a biased point of view, which can be seen in politics. In return of sharing one-sided views, the creator(s) intend the audience to adapt the same beliefs. Satire and parody are created for entertainment purposes more often than not. The Onion is an online 'news source' that publishes fake news stories in return for laughs. Sloppy journalism is, the majority of the time, published information without reliable sources and/or not fact checked. Misleading headlines are another tier of fake news but aren't entirely false. They are, however, easily distorted and easily misleading. Biased and slanted news is almost the opposite of propaganda. People already have certain set beliefs and will seek news and information on the same basis.

Previously, Americans would get the same content through each mass media outlet. Today, that's no longer the case, whether it's the content we want or the advertising. Social networking sites use algorithms to tailor newsfeeds to each of its users. It's a way of sorting posts on a person's feed based on components like relevancy, time and previous user activity;

more information on how algorithms work by using artificial intelligence (AI) will be defined later in this paper. Before the use of algorithms, most sites used the reverse chronological order method. In short, these algorithms determine what a user is seeing every time he or she logs on which in return can have a highly negative effect if what that user is seeing is fake news. This is why social media sites are attempting to combat the situation by removing content. The issue with this lay in the fact that fake news is not illegal, so, where's the First Amendment when this content is being forcefully removed?

Being involved in the distribution of fake news can have real—non virtual—life ramifications. In 2016, a former legislative aid to a Frederick County lawmaker in Maryland was fired after a fake political news website, Christian Times, he created was linked back to him. Cameron Harris, 23 at the time of the site, was responsible for spreading false information during the 2016 U.S. Presidential election. Harris profited approximately $22,000 in ad revenue from the site (Davis, 2019). He published well-fabricated articles about topics like pre-marked ballots to vote for Hillary Clinton that included images and sources to make it appear as if it was a true news article. These articles, as put by State Delegate David E. Vogt III, whom he worked for, promoted dishonesty and that was the grounds for Harris' removal from his legislative aid position (Davis, 2019). Fake news flourished during the 2016 U.S Presidential election because it supported bias news. People who were against Hilary Clinton flocked to articles like Harris' because it confirmed their own beliefs, no matter if those beliefs were right or wrong.

Another ramification the distribution of fake news can have is public panic. During the 2020 global pandemic of COVID-19, fake news about the virus spread rapidly on social media causing a public health crisis where people did not adhere to safety measures that ended up killing thousands of people and hospitalizing thousands more.

A major conspiracy that circulated was that President Trump would invoke the Stafford Act which would let him use martial law to enforce a mandatory shut down, or quarantine, of the American people. Multiple times during the spread of this story, the information was accredited to either a family friend or a friend of a friend who works for the government. The National Security Council debunked the rumor in two tweets that read, “Text message rumors of a national #quarantine are FAKE. There is no national lockdown. @CDCgov has and will continue to post the latest guidance on #COVID19. #coronavirus,” and “As we saw over the wkend, disinfo is being spread online about a supposed national lockdown and grounding flights. Be skeptical of rumors. Make sure you’re getting info from legitimate sources. The @WhiteHouse is holding daily briefings and @cdcgov is providing the latest” (National Security Council, 2020). Despite the tweets, the rumor was still spread.

Twitter, alongside a joint effort of other social media platforms, upped its safety policy due to the large distribution of misinformation surrounding COVID-19. On March 18, 2020, Twitter released a thread of tweets explaining how content that could place individuals at a higher risk of transmitting the virus would be removed that read, “Content that increases the chance that someone contracts or transmits the virus, including:

- Denial of expert guidance

- Encouragement to use fake or ineffective treatments, preventions, and diagnostic techniques

- Misleading content purporting to be from experts or authorities” (Twitter Safety, 2020).

A more detailed approach of what all of this meant was taken on their blog. The company listed eight steps that they would be taking in light of the global pandemic. One of those steps being is to increase the use of machine learning and automation. In a blurb explaining the

reasoning for an increase of AI, Twitter admitted that AI is not perfect and that mistakes would be made, “while we work to ensure our systems are consistent, they can sometimes lack the context that our teams bring, and this may result in us making mistakes” (Gadde & Derella, 2020). This meaning, censorship will be inevitable, and content will not be brought back on the site unless a complaint is made by the user.

Another step is to broaden the definition of harm on Twitter. In this step multiple areas were listed but two specifics were that any content that claims specific groups are not susceptible to COVID-19 would be removed as well as that, “Denial of established scientific facts about transmission during the incubation period or transmission guidance from global and local health authorities, such as “COVID-19 does not infect children because we haven’t seen any cases of children being sick.”” (Gadde & Derella, 2020). This, however, ended up being controversial after billionaire entrepreneur Elon Musk tweeted that, “Kids are essentially immune, but elderly with existing conditions are vulnerable. Family gatherings with close contact between kids & grandparents probably most risky” (Musk, 2020). Twitter has denied removing Musk’s tweet even though the statement that “kids are essentially immune” is false. In an interview with The Verge, Twitter said in a statement that, “When reviewing the overall context and conclusion of the Tweet, it does not break our rules. We’ll continue to consult with trusted partners such as health authorities to identify content that is most harmful” (O'Kane, 2020). Currently, the tweet is still online.

Facebook believes that, “False news is harmful to our community, it makes the world less informed, and it erodes trust,” which explains why the company is removing such content (Mosseri, 2017). In its attempts to fight this spread, Facebook is reducing the distribution of fake news by identifying which content is fake by using feedback from Facebook users as well as

third-party fact checkers and by removing repeat offenders. The fact checkers are used in certain countries who are certified through the non-partisan International Fact-Checking Network (Facebook, "How is Facebook addressing "). Facebook will also provide more information if there is fake news in a user's feed. When a fact-checker writes an article giving more context on the story, below the fake news article will be a 'Related Article' giving context on the fake news. In addition, if the user attempts to reshare the tainted information, he or she will get a notification that the post has been flagged by a fact-checker (Facebook, "How is Facebook addressing "). Lastly, Facebook wants to inform its users on how to spot fake news and will give the necessary tools to provide feedback on content they may believe is false.

Prior to the implemented policies mentioned above, during the 2016 presidential election, a study found that Facebook was the biggest culprit when it came to spreading fake news due to it having the largest audience of all social networks; bigger than Twitter, Google and all other webmail providers. In the journal *Nature: Human Behavior*, researchers led by Andrew Guess of Princeton University monitored over 3000 Americans and their internet use leading up to the 2016 election. Guess and his team discovered that Facebook was the referrer to fake news/untrustworthy sites over 15% of the time compared to its 6% of authoritative referrals (Guess, Nyhan, & Reifler, 2020). Whereas sites like Google accounted for 3.3% untrustworthy news and 6.2% authoritative news and Twitter for 1% untrustworthy news and 1.5% authoritative news (Guess, Nyhan, & Reifler, 2020).

Twitter's attempt to put an end to misleading information on the site was announced in 2019 when CEO Jack Dorsey published that they will no longer allow political advertisements. Political content is content either from or referencing, "a candidate, political party, elected or

appointed government official, election, referendum, ballot measure, legislation, regulation, directive, or judicial outcome" (Twitter, "Political Content").

While faced with backlash, Dorsey explained the reasoning in a tweet stating, "It's not credible for us to say: "We're working hard to stop people from gaming our system to spread misleading info, buuut if someone pays us to target and force people to see their political ad…well… they can say whatever they want!" (Dorsey, "For instance:", 2019). This tweet in particular was a subtle drag to Mark Zuckerberg, CEO of Facebook, after his decision to not block political speech on Facebook, no matter if it contains anything misleading. Zuckerberg told the U.S. House committee, "Our policy is that we do not fact-check politicians' speech, and the reason for that is that we believe that in a democracy, it is important that people can see for themselves what politicians are saying" (Schneider & Bond, 2019). Dorsey explains that this isn't a matter of free expression but rather a look at a new approach that regulators need to, "think past the present day to ensure a level playing field" (Dorsey, "In addition:", 2019).

Shadow Banning

Shadow banning occurs when a user's content is blocked without the user being aware. Based on complaints from users, this tactic is widely seen on Instagram; however, Instagram has not confirmed they use shadow banning. In fact, Instagram CEO Adam Mosseri shot down rumors in early 2020 when he stated, "Shadow banning is not a thing. If someone follows you on Instagram, your photos and videos can show up in their feed if they keep using their feed. Being in [Instagram's Explore page] is not guaranteed for anyone. Sometimes you'll get lucky, sometimes you won't" (Cook, 2020). However, Instagram does in fact partake in shadow banning. While the site will not specifically use the term shadow banning, it will use the terms hide or restrict when it comes to censoring certain posts from showing on the explore page—a

page consisting of photos and videos from accounts a user does not follow—or when searching a hashtag.

In the summer of 2019, Instagram was caught shadow banning posts from pole dancers and other attendees of a Caribbean carnival who had content with hashtags such as #PoleFitness, #StLuciaCarnival and #TrinidadCarnival2020 amongst a few others. This was proven when users clicked on the hashtags mentioned, they were prompted with a message that read, "Recent posts from [hashtag] are currently hidden because the community has reported some content may not meet Instagram's community guidelines" (Taylor, 2019). After receiving multiple complaints about the issue, an Instagram spokesperson released a statement that read, "The hashtags #xuvocarnival and #trinidadcarnival2020, among others, were restricted in error and we are restoring them to full visibility. We apologize for the mistake. Over a billion people use Instagram every month and operating at that size means mistakes are made—it is never our intention to silence members of our community" (Taylor, 2019).

Even though Instagram stated they do not shadow ban, they have acknowledged that the company does hide posts deemed inappropriate from the explore and hashtags pages—which is the definition of shadow banning to many individuals. The site uses a combination of automated filters and user reports to, "reduc[e] the spread of posts that are inappropriate but do not go against Instagram's Community Guidelines" (Constine, 2019). This can be described as borderline content. While a user's image may not contain nudity, which violates the Community Guidelines, it may be sexually suggestive which can be deemed inappropriate and therefore Instagram will hide it from other user's explore and hashtag pages.

This type of censoring is fairly new. In 2018, President Donald Trump signed the bill Allow States and Victims to Fight Online Sex Trafficking Act (FOSTA) to hold online platforms

responsible if there is content promoting or facilitating prostitution and/or sex trafficking. Since the passage of that bill, social media sites have adjusted their policies accordingly to FOSTA; however, the policies implemented leave room for purposeless censorship that does not violate FOSTA. Instagram demotes content deemed inappropriate, but the term inappropriate is up to interpretation. Facebook no longer allows sexually explicit language or content that "facilitates, encourages or coordinates sexual encounters between adults" (Facebook, "Community Standards"). Tumblr restricts "adult content" which they define as, "photos, videos, or GIFs that show real-life human genitals or female-presenting nipples, and any content—including photos, videos, GIFs and illustrations—that depicts sex acts" (Tumblr, "Adult content").

Marginalized and vulnerable communities such as women, LBTQ+ and minorities are being censored because of the removal and/or shadow banning of all and any content that fall under the guidelines aforementioned, not just ones that are in violation of FOSTA. The shift towards using automated filters to remove content grants an even greater confusion. How will automated technology be able to differentiate between a woman in lingerie and a woman in a bikini?

In the summer of 2019, Instagram mistook ads featuring LGBTQ people for escort ads. The issue? They were not escort ads; they were two trans women who were standing fully clothed. The ads were released by Salty, a newsletter and digital publication for women, transgender, gender non-conforming and non-binary people. Salty tweeted out two more ads that were not approved, both accompanied by the same escort ad message. The second image was of a disabled person-of-color on the telephone, fully clothed. The third image was of a person-of-color intersex activist laying on the ground, fully clothed, with the words "Love Our Hearts, Not Our Parts" laid out.

Salty received a notice from Instagram that the ads were not approved with a message that read, “The ads can’t be run because it doesn’t follow our Advertising Policies. We don’t allow ads that promote escort services. Ads like these are sensitive in nature and usually violate local laws, rules, or regulations” (Salty, 2019).

According to the publisher of Salty in an interview with Rolling Stone, she attempted to post the ads a week or so later, only to be brought with the same escort rejection message even after attempted contacts with Instagram to explain the people in the ads were not escorts (Dickson, 2019). After being denied multiple times, she brought the issue to Twitter, calling out both Facebook and Instagram. It was then when an Instagram spokesperson released a statement saying, “Every week, we review thousands of ads — and at times we make mistakes. We made mistakes here, and we apologize to Salty. We have reinstated the ads, and will continue to investigate this case to prevent it from happening again” (Dickson, 2019).

Artificial Intelligence

Artificial intelligence (AI) can be different to anyone who may utilize it, but at its core, it’s teaching machines to be and think like humans. There are two broad types of AI which are narrow AI and artificial general intelligence (AGI). Narrow AI is used every day. Smart phones use narrow AI for voice assistants and GPS navigation. Gmail uses narrow AI to automatically try to write and finish parts of an email. Services like Amazon, Netflix, or Hulu use narrow AI to recommend products it thinks its users may like. Social media sites like Instagram (as well as their parent company, Facebook)**,** as cited previously in this paper, are transitioning into using narrow AI for algorithms that create a user’s feeds as well as remove content that is either borderline or goes against the site’s community guidelines. Narrow AI, essentially, is the future of all social media platforms. Much of narrow AI is powered through machine learning. Machine

learning gives a computer data and uses statistical techniques that help the machine learn how to do the task better without having to a human program the machine through written code ("What is Artificial Intelligence? How Does AI Work?: Built In").

Using AI for algorithmic sorting is not shadow banning nor is it targeted censorship. The purpose for this strategy is to maximize the time users spend on the site by ensuring they see content that they are interested in. AI becomes a problem when its used to automatically hide or remove content based on certain criteria like sexual language. This is an issue because as mentioned before, how does AI pick up on the difference between a woman in lingerie which can be seen as sexually suggestive and therefore deemed inappropriate versus a woman in a bikini which does not violate any standards? This goes for AI and censoring language. How can AI understand context? If the president were to use profanity, would that be removed even though it is relevant?

Facebook uses ranking signals in order to curate a feed using AI. Ranking signals can be better defined as data points about a particular user's behavior as well as the behavior of all other users. The three main ranking signals are who a user typically interacts with, the type of media the post contains (e.g., video, text, link, image) and lastly, the popularity of the post (Cooper, 2020). For example, if a person is often interacting with their significant other on the platform, the significant other's posts are more likely to show first in the feed. Also, if a post is being shared in large amounts (i.e., in the thousands) it's likely to show in a higher position. To better develop its AI, Facebook in 2019 released a survey to all users that contained four questions. These questions were:

1. Who their close friends are;
2. What posts (links, photos and videos) they find valuable;

3. How important a specific Facebook Group that they've joined is to them;
4. How interested they are in seeing content from specific Pages that they follow.

All the questions were then used to update the newsfeed algorithm to better tailor to each individual user (Cooper, 2020).

However, AI still lacks the ability to assess emotion—both of the content creator and the viewer. The inability to understand emotion can play a role in how the content is perceived, consumed and appreciated.

AI is not perfect and therefore censorship is inevitable. In March 2020 during the global pandemic of COVID-19, Facebook was blocking users from posting legitimate news articles about COVID-19 because of a malfunction in its AI. The company had sent home many of its content moderators which lead to more reliance on AI because of the highly contagious factor of the virus. A former Facebook security executive speculated the aforementioned in a Tweet saying, "It looks like an anti-spam rule at FB is going haywire. Facebook sent home content moderators yesterday, who generally can't WFH" — work from home — "due to privacy commitments the company has made. We might be seeing the start of the ML" — machine learning — "going nuts with less human oversight" (Price, 2020). Despite the obvious shift in the content moderator workforce, Facebook denied the allegation that that factor had anything to do with the issue. The platform's vice president of integrity, Guy Rosen, tweeted, "We're on this – this is a bug in an anti-spam system, unrelated to any changes in our content moderator workforce. We're in the process of fixing and bringing all these posts back" (Price, 2020).

Like its parent company, Instagram uses ranking signals to curate newsfeeds. The three main components are similar to Facebook; it consists of relationship, interest and timeliness. Other factors that influence the order of posts are how often a person uses the app, the amount of

accounts a person follows and how long a person stays on the app (Cooper, 2020). Relationship, or how often a user interacts with another specific user is a main driving factor in where a post is ranked. The previous model of reverse-chronological order had users missing approximately 70 percent of their overall feed whereas the new AI ranking system allows users to see approximately 90 percent of posts from "family and friends" (Cooper, 2020).

In late 2019 and early of 2020, Instagram began removing posts and accounts that supported the killing of Iranian commander Qasem Soleimani to comply with U.S. sanction laws. News agencies, human rights activists and influencers were all having their accounts deactivated. Instagram is one of the few social media companies that are still accessible in Iran without having to use a virtual private network, making it one of the only platforms where Iranians can freely express themselves (Zakrzewski 2020). Iranian journalist Emadeddin Baghi, who is a critic of Iran's government, wrote an article about the killing of Soleimani stating it was "contrary to the principles of international law," shared it to Instagram to which it then was removed. This post did and does not go against the platform's policies but was still taken down. After facing backlash, Instagram restored the content and apologized, identifying an error that several posts were removed even though they do not violate any policies (Zakrzewski 2020). The question remains of how much content is being unrightfully removed due to AI?

Twitter curates its feeds a little different than Facebook and Instagram. It's a mix of both algorithm and real-time content (i.e., reverse chronological order); however, the majority of the feed is arranged in reverse chronological order. A user's timeline is organized into three main sections.

The first section of content is posts ranked by algorithm. It's important to note that there is no heading separating this content but if the user looks at the timestamps on the Tweets, its

evident that they're times in no particular order. There are four ranking signals that determine the algorithmic section of the timeline which are recency, engagement, rich media and activity (Nemeth 2020). Recency is equivalent to timeliness; how recent the Tweet is. Engagement deals with how many other users are interacting with that Tweet in the form of clicks, favorites and retweets. Rich media is what the content consists of, whether its images, videos or GIFs. Lastly, activity is how often the user of the Tweet is active. It's possible to see Tweets in this section from users that the account holder does not follow but has been engaged with by users he or she does follow. The second section is labeled, unlike the algorithmic section, "In Case You Missed It." Featured here are older Tweets based on the algorithm, however this section is not always seen every time a user views their timeline and it only consists of a small number of Tweets. The last section is the traditional reverse-chronological order.

Twitter wants its users to have healthy conversations and in order to do that, they're using machine learning. The platform removes any accounts that goes against their policies, but the issue is with accounts that aren't going against policies but are creating borderline content which in return is "distor[ing] the conversation" (Harvey & Gasca 2018). In a blog post, the company wrote, "To put this in context, fewer than 1% of accounts make up the majority of accounts reported for abuse, but a lot of what's reported does not violate our rules. While still a small overall number, these accounts have a disproportionately large – and negative – impact on people's experience on Twitter. The challenge for us has been: how can we proactively address these disruptive behaviors that do not violate our policies but negatively impact the health of the conversation?" (Harvey & Gasca 2018).

Law

What does it mean for the First Amendment when people working in politics, either at a state or federal level, are using social media for political expression? Though the First Amendment generally has no legal value in private companies, it did in Knight Institute v. Trump. The Knight Institute is a foundation to protect free expression during the digital age. This was a lawsuit about President Trump blocking critics on his personal Twitter account and whether or not this was legal due to it being on Twitter, a private company, and his presidential status. The United States District Court of the Southern District of New York concluded that when President Trump blocks a person on Twitter, he is violating the First Amendment. This is because when doing so, he is discriminating against other's viewpoints and that prevents them from both participating in debates on Trump's page as well as be informed when Trump tweets government related news ("Knight First Amendment Institute v. Donald J. Trump", 2019). In simpler terms, the president's communication is a public forum and therefore cannot be limited to anyone.

This case has since been appealed and in its appeal the verdict did not change. In the Knight v. Trump amicus brief for the United States Court of Appeals for the Second Circuit it reads, "Given the pervasive use of social media, this Court must recognize that individuals have First Amendment rights both to receive governmental messages transmitted through social media as well as to participate in the interactive communicative forums created by them. And this Court must find that the President's viewpoint-based blocking of the plaintiffs burdens their First Amendment rights, and is thus unconstitutional" ("amicus brief", 2018).

Knight v. Trump began the conversation at a legal level for the First Amendment and social media. In January of 2019 the first federal appellate court in the verdict of Davison v. Randall decided that government officials cannot dictate what views appear on government

social media pages. Loudoun County School Board (LCSB), in Virginia, Chair Phyllis Randall created a Facebook page for her office, which is considered a public forum under the First Amendment after the verdict in Knight v. Trump, and deleted comments of a critic ("Davison v. Randall", 2019). Brian Davison, the critic, had commented allegations of corruption and conflicts of interest on Randall's page and in return, Randall deleted the comments and blocked Davison from the page for approximately 12 hours. During this time, Davison was unable to see the page or engage with it in any kind of manner. He believed that this was a violation of his free speech and due process rights under the U.S. and Virginia state constitutions ("Davison v. Randall", 2019). The US Court of Appeals for the Fourth Circuit agreed with the decision of the District Court in that Randall acted in an unlawful manner and violated Davison's First Amendment rights.

Not long after Knight v. Trump in April of 2019, the Fifth Circuit also agreed that government officials cannot delete nor block comments on social media in Robinson v. Hunt County. The Hunt County Sheriff's Office (HCSO) has a Facebook page and during the time of this case under their "About" section, it read, "Welcome to the official Hunt County Sheriff's Office Facebook page. We welcome your input and POSITIVE comments regarding the Hunt County Sheriff's Office. The purpose of this site is to present matters of public interest within Hunt County, Texas. We encourage you to submit comments, but please note that this is NOT a public forum" (Justia, "Robinson v. Hunt County,").

On January 18, 2017, the HCSO posted an update that read, "We find it suspicious that the day after a North Texas Police Officer is murdered we have received several anti-police calls in the office as well as people trying to degrade or insult police officers on this page. ANY post filled with foul language, hate speech of all types and comments that are considered

inappropriate will be removed and the user banned. There are a lot of families on this page and it is for everyone and therefore we monitor it extremely closely. Thank you for your understanding" (Justia, "Robinson v. Hunt County,"). Due to the HCSO's understanding that the page is not considered a public forum, they believed they had the right to delete any comments or posts that violated the rules outlined in the "About" section. However, Deanna Robinson, who sued the HCSO office, alleged that the HCSO Facebook page is a public forum and therefore the office does not have the right to delete any comments. Robinson, amongst other users, had commented on the post mentioned above with criticism that by HCSO deleting comments, it is censorship. According to Robinson, soon after she commented the allegations about free speech and censorship the HCSO office removed her comments and banned her from the page (Justia, "Robinson v. Hunt County,"). The Fifth Circuit found that deleting Robinson's comment was unlawful and discriminatory, however, the court did not classify what type of forum the HCSO created.

Just like Knight v. Trump and Robinson v. Hunt County, the American Civil Liberties Union (ACLU) in 2017 sued Maryland Gov. Larry Hogan and two of his aides for blocking Facebook users after they had left comments on his Facebook page. The Facebook users who were blocked and represented by the ACLU, argued that Hogan was censoring their free speech after they had left comments questioning both his position and education policy as well as President Trump's travel ban (Wiggins 2017).

According to Hogan spokesman Doug Mayer, between 2015 and 2017 Hogan blocked 450 people from his Facebook page. Mayer said that approximately half of those users were blocked for using hateful or racist language and the other half after the 2014 riots in Baltimore (Wiggins & Nirappil 2017).Under Hogan's original social media policy, comments would be

removed if it contained profanity or obscenities, a hyperlink to inappropriate content, was repetitive to other user's comments, was not posted regarding the content in the post, or the user would be blocked if he or she threatened violence or participated in a coordinated effort (Wiggins 2018).

In 2018, several of the Facebook plaintiffs under ACLU's council, reached a settlement in the case. Approved by the state Board of Public Works, the state had to pay $65,000 to the plaintiffs and attorney fees as well as develop a new social media policy for Hogan, create a separate Facebook page for users to discuss and raise issues and set up an appeals process for other users who may believe their comments were unrightfully removed (Wiggins 2018).

While the Supreme Court has established that government officials cannot censor on a private social media platform, this does not apply to private individuals. Private social media sites have no obligation to abide by the First Amendment to non-government officials which was seen in the court case Nyabwa v. Facebook.

In 2016 Collins O. Nyabwa created a website called "emolumentsclause.com" to inform others about business conflicts of interest with President Trump. Nyabwa registered a Twitter account under the same name as his website and attempted to do so with Facebook. Facebook locked the account within a few days of its creation with a message reading, "For security reasons your account is temporarily locked" (Hayden, 2018). The network asked Nyabwa to upload a government issued photo ID to prove his identify to unlock the account, which Nyabwa did. However, Facebook did not unlock the account, leading to Nyabwa to form a complaint explaining that the account still being locked after providing a valid government ID is unconstitutional. He believed that Facebook was not worried about his personal identity, but rather did not agree with his political beliefs (Hayden, 2018). The court ruled that a private

individual cannot sue for free speech against Facebook, writing “the First Amendment governs only governmental limitations on speech” (Hayden, 2018).

Conclusions

Social media censorship is a public concern and therefore people deserve a public answer. The way social media is used as means of communication makes it seem as if one’s First Amendment rights should be protected. With fake news, shadow banning and artificial intelligence, many users are confused as to what they can and cannot do. Since platforms like Twitter, Facebook and Instagram are private, they can police and censor as they please, but since the majority of public discussion takes place on these platforms, it’s become a concern for freedom of speech.

Artificial intelligence is the future of both social media and everyday life; it’s not going away. The only way to make AI better, is to expose it to more situations to learn from. That’s where the American people are today and where some Americans are having issues adjusting to. Someone’s post got taken down because she was in a bikini, but now that AI has experienced a new situation it will learn that content is appropriate and does not violate any policies. Machine learning needs exposure and the only way to get that is to accept that some mistakes will be made. It’s essential that these platforms are upfront about their AI uses and that some mistakes will be inevitable. With clear policies, users will be more open to understand and to bear with the bumps in the road.

Shadow banning is usually the fault of AI. Since AI is so commonly used, it’s important for companies to have non-virtual workers monitoring what AI is taking down and to be open to feedback from its users. Overall, social networking sites only work when there’s user-generated

content. If the users aren't happy, there won't be any content and there won't be a sustainable business.

Fake news, while nothing new, is a real issue. As seen with the global pandemic of COVID-19, there are real-life ramifications that the spread of misinformation can have. Though these social media platforms are attempting to remove all kinds of misinformation, there are errors and complaints from users that they believe their First Amendment rights are being infringed upon. To keep users happy, since there is no legal reason to follow the First amendment for these companies, sites like Facebook try to keep the content up and inform users what is fake and give them options to either remove it from their newsfeed or read the authoritative version.

As the model of communication continually adapts to new technology, it is important to examine the areas mentioned above alongside the First Amendment and whether or not private corporations should have a legal obligation to follow.

References

Constine, J. (2019, April 10). Instagram now demotes vaguely 'inappropriate' content. Retrieved from https://techcrunch.com/2019/04/10/instagram-borderline/

Cook, J. (2020, February 27). Instagram's CEO Says Shadow Banning 'Is Not A Thing.' That's Not True. Retrieved from https://www.huffpost.com/entry/instagram-shadow-banning-is-real_n_5e555175c5b63b9c9ce434b0

Cooper, P. (2020, February 5). How the Facebook Algorithm Works in 2020 and How to Work With It. Retrieved from https://blog.hootsuite.com/facebook-algorithm/

Cooper, P. (2020, April 20). How the Instagram Algorithm Works in 2020 (And How to Work With It). Retrieved from https://blog.hootsuite.com/instagram-algorithm/

Davis, P. (2019, June 22). Legislative aide in Annapolis fired after Times reveals him as fake news mastermind. Retrieved from https://www.capitalgazette.com/politics/ph-ac-cn-fake-news-annapolis-0120-20170119-story.html

Davison v. Randall. (2019). Retrieved from https://globalfreedomofexpression.columbia.edu/cases/davison-v-randall/

Dickson, E. J. (2019, July 11). Why Did Instagram Confuse These Ads Featuring LGBTQ People for Escort Ads? Retrieved from https://www.rollingstone.com/culture/culture-features/instagram-transgender-sex-workers-857667/

Dictionary.com. (n.d.). Obscene. Retrieved from https://www.dictionary.com/browse/obscene

Dorsey, J. (2019, October 30). In addition: Retrieved from https://twitter.com/jack/status/1189634374758617088?s=20

Dorsey, J. (2019, October 30). For instance: Retrieved from https://twitter.com/jack/status/1189634371407380480?s=20

Facebook. (n.d.). Community Standards: Sexual Solicitation. Retrieved from https://www.facebook.com/communitystandards/sexual_solicitation

Facebook. (n.d.). How is Facebook addressing false news through third-party fact-checkers?: Facebook Help Center. Retrieved from https://www.facebook.com/help/1952307158131536

Facebook. (n.d.). Terms of Service. Retrieved from https://www.facebook.com/terms.php

Facebook. (n.d.). Community Standards. Retrieved from https://www.facebook.com/communitystandards/

Gadde, V., & Derella, M. (2020, April 1). An update on our continuity strategy during COVID-19. Retrieved from https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.html

Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*. doi: 10.1038/s41562-020-0833-x

Harvey, D., & Gasca, D. (2018, May 15). Serving healthy conversation. Retrieved from https://blog.twitter.com/official/en_us/topics/product/2018/Serving_Healthy_Conversation.html

Hayden. (2018, January 26). Nyabwa v. Facebook. Retrieved from https://casetext.com/case/nyabwa-v-facebook/?PHONE_NUMBER_GROUP=P&NEW_CASE_PAGE=N

Jr, D. L. H. (2019, April 1). Free speech or censorship? Social media litigation is a hot legal battleground. Retrieved from https://www.abajournal.com/magazine/article/social-clashes-digital-free-speech

Justia. (n.d.). Robinson v. Hunt County, No. 18-10238 (5th Cir. 2019). Retrieved from https://law.justia.com/cases/federal/appellate-courts/ca5/18-10238/18-10238-2019-04-15.html

Knight First Amendment Institute v. Donald J. Trump. (2019). Retrieved from https://globalfreedomofexpression.columbia.edu/cases/knight-first-amendment-institute-v-donald-j-trump-2/

Knight v. Trump amicus brief. (2018, October 19). Retrieved from https://www.eff.org/document/knight-v-trump-amicus-brief

Mosseri, A. (2017, April 2). Working to Stop Misinformation and False News. Retrieved from https://www.facebook.com/facebookmedia/blog/working-to-stop-misinformation-and-false-news

Musk, E. (2020, March 19). Kids are essentially immune. Retrieved from https://twitter.com/elonmusk/status/1240758710646878208

National Security Council. (2020, March 20). As we saw. Retrieved from https://twitter.com/WHNSC/status/1240808335349342208

Nemeth, C. (2020, March 12). How the Twitter algorithm works in 2020. Retrieved from https://sproutsocial.com/insights/twitter-algorithm/

O'Kane, S. (2020, March 20). Twitter won't remove irresponsible Elon Musk tweet about coronavirus. Retrieved from https://www.theverge.com/2020/3/20/21187760/twitter-elon-musk-tweet-coronavirus-misinformation

Perrin, A., & Anderson, M. (2019, April 10). Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018. Retrieved from

https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/

PEW. (2019, June 12). Demographics of Social Media Users and Adoption in the United States. Retrieved from https://www.pewresearch.org/internet/fact-sheet/social-media/

Price, R. (2020, March 17). Facebook is wrongly blocking news articles about the coronavirus pandemic. Retrieved from https://www.businessinsider.com/facebook-blocking-coronavirus-articles-bug-2020-3?fbclid=IwAR06brbzO-b4nBH8LhO3iA_NV1O9fcSuskXLnkaSStTSxHMFgz3VDidpBNs

Salty. (2019, July 9). A thread: Retrieved from https://twitter.com/Saltyworldbabes/status/1148741480154042369?ref_src=twsrc^tfw|twcamp^tweetembed|twterm^1148741480154042369&ref_url=https://www.rollingstone.com/culture/culture-features/instagram-transgender-sex-workers-857667/

Schneider, A., & Bond, S. (2019, October 30). Twitter To Halt Political Ads, In Contrast To Facebook. Retrieved from https://www.npr.org/2019/10/30/774865522/twitter-to-halt-political-ads-in-contrast-to-facebook

Taylor, S. (2019, July 30). Instagram Apologizes For Blocking Caribbean Carnival Content. Retrieved from https://www.vice.com/en_ca/article/7xg5dd/instagram-apologizes-for-blocking-caribbean-carnival-content

Tumblr. (n.d.). Adult content. Retrieved from https://tumblr.zendesk.com/hc/en-us/articles/231885248-Sensitive-content

Twitter. (n.d.). Political Content. Retrieved from https://business.twitter.com/en/help/ads-policies/prohibited-content-policies/political-content.html

Twitter. (n.d.). Twitter Terms of Service. Retrieved from https://twitter.com/en/tos

Twitter Safety. (2020, March 18). Content that increases. Retrieved from https://twitter.com/TwitterSafety/status/1240418440982040579?ref_src=twsrc^tfw|twcamp^tweetembed|twterm^1240418440982040579&ref_url=https://techcrunch.com/2020/03/18/twitter-coronavirus-covid-19-misinformation-policy/

Web Wise. (2020, March 2). Explained: What is Fake news?: Social Media and Filter Bubbles. Retrieved from https://www.webwise.ie/teachers/what-is-fake-news/

What is Artificial Intelligence? How Does AI Work?: Built In. (n.d.). Retrieved from https://builtin.com/artificial-intelligence

Wiggins, O., & Nirappil, F. (2017, February 8). Gov. Hogan's office has blocked 450 people from his Facebook page in two years. Retrieved from https://www.washingtonpost.com/local/md-politics/gov-hogans-office-has-blocked-450-people-from-his-facebook-page-in-two-years/2017/02/08/54a62e66-ed45-11e6-9973-c5efb7ccfb0d_story.html

Wiggins, O. (2017, August 1). Gov. Larry Hogan sued by ACLU for deleting comments, blocking Facebook users. Retrieved from https://www.washingtonpost.com/local/md-politics/md-aclu-sues-governor-for-deleting-comments-and-blocking-facebook-users/2017/08/01/9723d4a6-76d8-11e7-9eac-d56bd5568db8_story.html

Wiggins, O. (2018, April 2). Maryland, ACLU reach settlement over governor deleting critical comments on his Facebook page. Retrieved from https://www.washingtonpost.com/local/md-politics/maryland-aclu-reach-settlement-over-governor-deleting-critical-comments-on-his-facebook-page/2018/04/02/8b3073a4-3684-11e8-8fd2-49fe3c675a89_story.html

WIPO. (n.d.). Types of intellectual property. Retrieved from https://www.wipo.int/about-ip/en/

Zakrzewski, C. (2020, January 13). The Technology 202: Instagram faces backlash for removing posts supporting Soleimani. Retrieved from https://www.washingtonpost.com/news/powerpost/paloma/the-technology-202/2020/01/13/the-technology-202-instagram-faces-backlash-for-removing-posts-praising-soleimani/5e1b7f1788e0fa2262dcbc72/