Social Media As a Public Good: Finding a Compromise Between Public And Private Interests In Social Media Content Moderation

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Joseph Spaeth
Spring 2021

On my Honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assingments

Advisor
Kathryn A. Neeley, Associate Professor of STS, Department of Engineering and Society

Introduction

The spread of misinformation and hate speech online causes harm not only for end users, but also social media companies, reputable news sources, and broadly, discourse as a whole. The ability for false news and hate speech to propagate quickly has defined the last 5 years and will continue to be an issue going forward. The consensus is that this is caused by human actors who share posts not based upon the reliability of the content but instead based on confirmation bias, the tendency for people to believe things they already agree with regardless of whether or not that information is true (Dizikes, 2018; Nguyen 2019). Sites have taken steps to moderate their content and harassment to varying degrees of success, from Tumblr's automated pornography detection system, or Facebook's hate speech recognition system (Gillespie 2020 3, Thrasher 2018). As social media continues to replace traditional news sources and as it continues to grow in size, it is vital that social media sites implement clear and consistent content moderation policies so as to avoid the continued spread of problematic content on these platforms..

Currently the question of how to moderate social media content is of the utmost importance and they are still exploring the different possibilities for content moderation. This in conjunction with the results of my technical work in identifying echo chambers (insular online communities) puts us in a position where determining better content moderation is key, as it will allow for decisions with regards to bans, suspensions, and censorship to potentially be made earlier and more aggressively. This in conjunction with the uncertainty of how social media sites handle content moderation create the issues we see on websites like Twitter and Facebook currently. We also face free speech concerns with regard to these issues. How can one respect free speech without allowing bad actors to poison the discourse? As we continue to gather more data from these sites, as well as iterate on methods of analyzing that data (such as my technical work) it will become increasingly vital for social media sites to moderate their content, and moderate it well.

Should content moderation continue in its current forms on popular social media sites, they risk pushing users to unmoderated platforms where echo chamber formation is actively encouraged and worsened. Examples of these include the explicitly far-right platforms such as Gab and Parler (Zannetou 2018 1). Sites like these can remain problematic, as Zannetou claims "small 'fringe' Web communities within Reddit and 4chan can have a substantial impact on large mainstream Web communities like Twitter." (Zannetou, 2). Additionally, current content moderation techniques do not do a particularly good job of keeping problematic content off of their platforms. This can be seen in how Twitter and Facebook are handling the unsubstantiated claims of election fraud making their way to through the platforms (Kelion, 2020). Without improvements in content moderation similar misinformation will continue to dissipate through these platforms. Additionally, the question of whether the government should hold social media services accountable for the actions of their users needs to be explored, as it is currently a topic of much debate in US politics (Kelion 2020). Particularly relevant, is the question of repealing Section 230 of the Communications Decency Act (Communications Decency Act 1996).

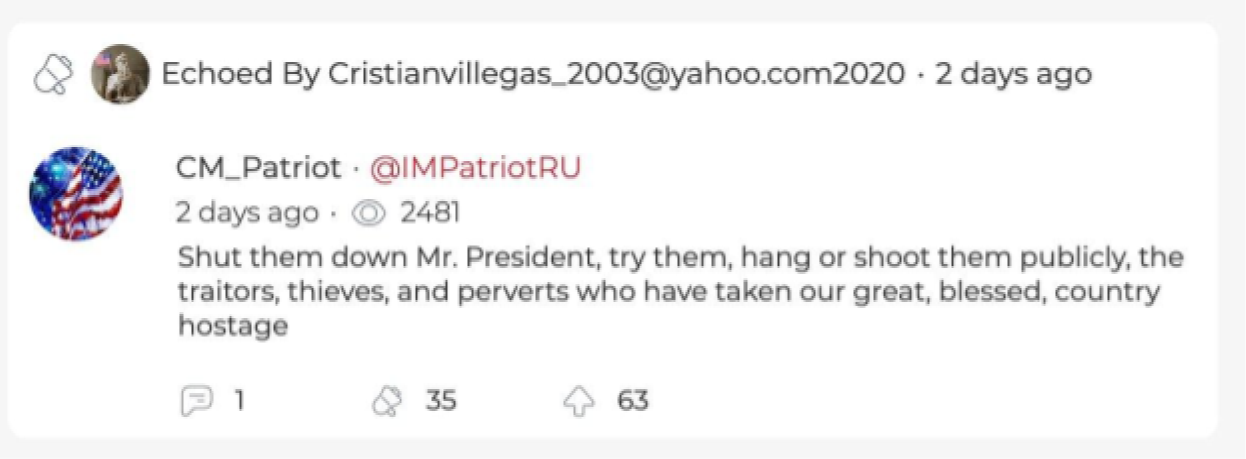
To assist in solving this problem, I propose this analysis of literature regarding content moderation on social media, as well as specific policy guidelines which can be (and in some cases have successfully been) implemented by social media sites. These policy proposals include increased use of techniques like minimizing the automation of content moderation, fact checking potentially misleading posts, and continuing with bans only for particularly egregious offenses. In this paper, I argue social media ought to be analyzed as a public good, that social media content moderation is a public policy issue, and I create a framework which policy proposals to handle content moderation should follow. This is done with the aforementioned literature review, followed by applying Mesthene's public and private interests framework to show social media is indeed a public good, and finally, a policy framework.

How Content Moderation Is Done Must Be Changed, Examining Some Case Studies

At the moment, we understand that people are the driving cause of the spread of misinformation and hate speech on the internet via social networks (Dizikes 2018, Nguyen 2019). Twitter and Facebook have been making progress in the field of content moderation, with bans of high profile users spreading disinformation, and laying out a plan to make it less economically feasible to spread false news, as well as a media literacy campaign respectively (Twitter, 2021 Facebook). However, this has not proven to be enough to keep false news and hate speech from the platforms. There are currently several different means of social media content moderation with varying degrees of acceptance and effectiveness. Popular content moderation methods include automated systems, user reporting, manual content moderation, user banning, shadow banning, demonetization, and the addition of fact checking flags, among others. These have been used to varying degrees of success, as well as varying degrees of appreciation from the end user and the platforms themselves. For example, some users are upset with Tumblr's automated porn detection system for being too aggressive and tagging things that are not porn, while others are upset with Facebook's automated hate speech detection for not tagging hate speech in comments and posts, as well as tagging things which are clearly not hate speech. Other measures have met more widespread approval, both from users and the companies themselves, such as Twitter adding flags on misleading posts, or Twitter and Facebook mentioning who is behind any political ad on the platforms. We also know that users will flee to alternative platforms with little to no moderation where echo chamber formation and the spread of misinformation and hate speech will become even worse if platforms are moderated too heavily. Here I examine the consequences of the content moderation policies used by popular sites and see the effects they have had.

Case Study 1: Website Bans, a Blunt and Ineffective Solution

For example, in the wake of the January 6th Capitol riot Donald Trump (as well as several other prominent conservatives) was permanently banned from Twitter for his role in inciting the violence and abuse of the platform (Twitter 2021). Some were surprised that he did not jump to Parler, a conservative "Twitter clone" marketing itself as a haven for free speech which had gained notoriety leading up to the presidential election that November. Parler fills a similar niche to Gab, with both operating as havens for the alt-right online. These platforms act as far-right echo chambers, and the speech that a website like Twitter would ban from its platform is allowed to spread freely here (Zannetou, 2). In fact, many users came to those sites after bans from other platforms. After the Capitol riot, Amazon Web Services (AWS) announced it would no longer host Parler and a researcher scraped the website in order to preserve its contents. It was filled with calls to violence, Figure 1 shows a particularly egregious example (Petkauskas 2020).



Echoed By Cristianvillegas_2003@yahoo.com2020 · 2 days ago

CM_Patriot · @IMPatriotRU

2 days ago · 2481

Shut them down Mr. President, try them, hang or shoot them publicly, the traitors, thieves, and perverts who have taken our great, blessed, country hostage

1 35 63

Figure 1: An example of hateful content and calls to violence from Parler (Petkauskas).

This was not an atypical post on the site, particularly in the days surrounding the violence. Harsh moderation methods like bans, though solving the problem on the platform one bans users from, can cause the formation of far worse communities outside of that platform.

Case Study 2: The Pitfalls of AI and Automation

We also have to consider the effects of using automated systems. Facebook, Youtube, and Tumblr all rely on automated content moderation to either remove or flag content. None of these systems work as intended, and in the cases of youtube and facebook, have actively hurt harmless online communities. Particularly, I examine Tumblr and Youtube. Tumblr explicitly banned pornography from its platform in 2018, using bots remove anything which violated the site's new Terms of Service (TOS). The effort was in order to prevent the spread of child pornography on the platform, but it led to the removal of all porn from Tumblr, as well as removal non-pornographic posts. Thrasher cites the example of a drawing of 2 men hugging being removed by the automated system (Thrasher 2018). Additionally, medical imagery is often removed, as is LGBTQ+ focused content. Similarly, Youtube relies on an algorithm to determine what is okay to monetize, what is not, and what Youtube will place ads on without paying the creators for. Youtube also has automated systems in place to prevent kids from seeing some videos. Similarly to Tumblr, this is a noble goal with a failed execution. Several LGTBQ+ content creators, particularly transgender creators, have seen their content be demonetized and age-locked despite being child friendly (Farokhmanesh, 2018). This harms Youtube's reputation, users, and creators on the platform. Ultimately, there is more nuance in content moderation decisions than what we can reliably expect computers to make. Though some sites have appeals policies, they are slow, and in the case of a site like Youtube, where money is on the line for content creators filing appeals, the money lost between when a video is flagged for age-restriction and when an appeal finishes can be massive.

Case Study 3: Demonetization to Disincentivize

Demonetization, the practice of preventing creators from making money off of their posts, is seen as a viable strategy for combating the spread of misinformation and hate speech online. However, it needs to be targeted to be effective, as demonstrated with the example of Youtube. Facebook has taken to demonetizing platforms which consistently publish false content (Facebook). Additionally, Facebook looks to be more surgical in their approach compared to a platform like Youtube. They are stopping ad buys from accounts associated with publishing false or misleading information, adding fact checking to posts (a policy decision also being taken by Twitter), and aggressively taking down spam accounts, and instead of using machine learning to find false news, they are using it to find bots publishing it (Facebook, Twitter, 2021). Though the platform still has issues with it, Facebook is not actively demonetizing accounts that do not deserve it, as well as ensuring that humans are more involved in the process compared to a website like Youtube (Facebook). They also take steps to ensure that content flagged as false or misleading is placed lower in the page rank algorithm, making it harder to see false news even when those stories are going unpromoted (Facebook).

There are several issues with the current content moderation paradigms of social media platforms. We lack an understanding of how current content moderation solutions affect stakeholders in the long term. We also do not have a complete understanding of what policies will push users to other platforms, beyond the fact that it is generally aggressive content moderation that does so. We also do not know what the effects of content moderation on echo chambers as they form will be. This is a more aggressive stance and could potentially end up backfiring depending on the methods used, as seen with users fleeing to platforms like Gab and Parler. There is a fine line that social media sites must walk between ensuring that users have access to free speech and the ability to have open discussions while also keeping the platforms safe and free of misinformation. The biggest uncertainties in the field of content moderation lie in determining what needs to be a ban-worthy offense on social media sites, the values of the sites themselves, how to ensure users do not congregate in other corners of the web allowing misinformation or hate speech to continue to propagate, and how to ensure that innocent users to do not get left by the wayside as a result of content moderation policies.

Method	Pros	Cons
Bans and Suspensions	Immediate, simple, prevents further spread of problematic content from a given user	Banned users will congregate in other online spaces with other like minded users
Automation	Minimal human interaction is required after implementation	AI is ineffective
Demonetization	Limits reach of problematic content, does not drive users off the platform	Creators can and will seek monetary opportunities through other platforms

Figure 2: Content moderation methods alongside their pros and cons

At this point, I have demonstrated what current content moderation efforts typically do, and have shown where their successes and shortcomings lie. Having an understanding of the options available to platforms helps provide knowledge of both what can be done, and what is currently viewed as best practice. With that in mind, we now turn to the question of who is ultimately responsible for providing those content moderation decisions, and what their responsibilities are.

Applying Mesthene's Perception of Technology and Society, and The Balance of Public and Private Good

In chapter 3 of his book "*Technological Change Its Impact on Man and Society*" Emmanuel Mesthene introduces his idea of technological advancement affecting larger social systems over time and framework of balance among private and public interests are vital to understanding the issue of online content moderation. He defines “political” broadly, as any decision that is relevant to public policy, regardless of who may be making those decisions (Mesthene 63-64). This definition, as well as those of other specialized terms defined by Mesthene, can be seen in Figure 3. In this chapter, Mesthene asserts two major things.

Term	Definition
Political	“All decision making structures and procedures that have to do with the allocation and distribution of power and wealth in society” These structures are both private and public.
Private entity	Any non-government entity. Mesthene cites “firms, labor unions, churches, political parties, professional or trade associations and individuals” as examples.
Public entity	Government entities, for example, Congress.

Figure 3: Mesthene’s definitions of technical terms (created by author). These are the definitions used by the author.

The first is that as technology advances and grows more ubiquitous, it goes from a private luxury to a public good. The second is that as that evolution happens, a balance needs to be struck between private and public interests in order to continue and manage further technological development. In essence, as the technology becomes more readily available, the decisions surrounding it will inherently become more political, as they affect far more people and the interactions between systems become more complex. Mesthene cites public schooling as an example of this. Initially, schools in the United States were few and far between, but as schooling became more important, more and more governmental structure needed to be integrated to ensure consistent curricula and that students had access to public schooling (Mesthene 65). I assert that Mesthene’s framework of balancing public and private interests in developing technology is applicable to the question of online content moderation because of social media’s rise to ubiquity in the last 15 years, and it acting as a de facto public commons online.

The reach of websites like Facebook and Twitter is massive, and prior to either of those websites existing the internet had already become a means of spreading large amounts of information. This led to the creation of Section 230 of the Communications Decency Act, the United States' Congress' first attempt to legislate content published online. The law does two things, it makes it so the government cannot treat a website as the publisher of its contents (for example Twitter is the publisher of any user's tweets) and it carves out protections for websites to moderate content both legal and illegal as well as encouraging those websites to do so (Communications Decency Act 1996). This marks the beginning of online spaces becoming a public good, and that technology entering the public sphere in the United States. This is also a demonstration of how over time technological developments become more integrated into public life. Alongside that integration, it is necessary for it to be regulated with public interests in mind. These things taken together demonstrate how cleanly Mesthene's framework maps onto the technology of the internet and social media. Section 230 explicitly encourages content moderation, in exchange for websites not being treated as publishers. It is an initial compromise between the United States Government and online platform owners, a means by which both share power to reach a public policy goal. This compromise demonstrates a straightforward application of Mesthene's ideas in the sphere of content moderation, and where to continue going forward.

Additionally, new attempts have come from the US government to regulate social media. This demonstrates the conditions discussed by Mesthene, the decision making with how those sites handle content moderation has become deeply political, and neither the government nor the companies controlling these sites are capable of adequately making these public policy decisions alone. Mesthene asserts that neither private interests nor the government are capable of making the ideal decisions with regard to this still developing technology. Instead, we need input from both, he says “With the proper economic and political organization, we could derive greater benefits from our technology than we do” and what he means by this is that public and private interests need to work closely together as this technology continues to develop (Mesthene, 74). This is because the government's best interests are the public’s (or at the very least, ought to be) and corporations have the familiarity with the technology to ensure a high quality product is delivered (while instead having profit be their primary interest). Accordingly, to develop a public policy framework, one must couple together the public and private sector, further evidence that this method is both applicable and relevant to the content moderation issue.

Mesthene’s idea of technology breaking into the public sphere, and framework of balance between public and private interests in society is vital to understanding social media moderation. Though almost all social media sites are privately owned, they have become the go to place for the dissemination of information on the internet. Social media and the internet more broadly are technologies which will not go away. As speech continues to centralize on the internet the political decision of what is allowed in the digital commons will have to be made by both governments and private companies, and a delicate balance must be struck to protect the interests of both.

Answering The Public Policy Question: How To Handle Content Moderation

As demonstrated in the prior section, it is helpful to think of social media as a public good, as the decisions made surrounding it are political. Social media companies have begun to step up and improve their content moderation efforts, and this question is being taken seriously in the American political sphere. As of late, the tone between legislators and the owners of Facebook and Twitter has been adversarial, and the law makers have yet to find common ground on what reforms (if any) to pass in the wake of the massive misinformation campaigns surrounding the 2020 elections (Kelion 2020). I however, am opting to view this issue from a place where the public and private work together. At this point the internet and even individual sites are so large that oversight cannot effectively come from solely the government or a private organization. This means a policy must be drafted which protects the following: a government interest in freedom of speech, the interest of social media companies in growing their platforms, the responsibility of those companies to their shareholders, and the safety and well being of the user base. I claim that social media sites and the states the operate out of must collaborate to create policy which:

1. Has a ree speech model similar to those of Western Europe and Canada, using people instead of computers to make moderation decisions
2. Provides safeguards against the spread of misinformation
3. Has clear communication of what policy violations are and their punishments

will be the ideal way to handle online content moderation.

Canada is much more aggressive in handling hate-speech than the United States is. Canada's criminal code has an outright ban on "wilful promotion of hatred", which exists to protect both individuals and "society at large" (Moon 85, 87). It is important to note however, that Canada's hate speech laws only apply to speech which could credibly breach the peace (Moon 87). This however, still protects most forms of hate speech and leaves the door open for racist acts. At this point, the duty of the government ends, and social media sites must pick up the slack. This would primarily come in the form of large scale human content moderators, due to the failures of AI in this area. Beyond the fact that Machine Learning is often ineffective in detecting hate speech, on Facebook, the vast majority of automatically tagged hate-speech was tagged by a human before any AI tagged it (Gillespie 3). A shift in the conception of free speech as well as an increase in the number of human moderators is the first step toward successfully moderating hate-speech and misinformation.

To combat false news we must look at other policies meant to curb the same thing, and how both the government and social media companies can step in to fill this role. Looking again to free speech laws, one can see that Holocaust Denial is not considered protected speech in many European countries, and that there is a compelling case to have it be treated as unprotected speech in the United States. In his article on "Socially Worthless Untruths" Steven Gey points out that there is a vested interest in limiting the dissemination of misinformation, as "Conducting discussions of public policy in a context where dissemination of false facts is permitted, on the other hand, provides nefarious political actors with the opportunity to convince uncommitted citizens that there is a factual basis for "bad" opinions that the uncommitted citizens would otherwise reject" (Gey 14). Similarly, Germany has several restrictions on racist speech and holocaust denial (as both are heavily tied to naziism) in an attempt to curb similarly "bad" opinions in Germany (Knechtle 49). There is an inherent government interest for the sake of political discourse in limiting the spread of misinformation, as otherwise it becomes significantly harder to have political conversations and for democracy to work. Particularly as the US government has continued to investigate Facebook and Twitter executives over the misinformation spreading on their platforms, it is becoming incredibly clear that content moderation is a very political (as Mesthene would use the word) issue. This creates a space where the interests of both the social media platforms, and government line up, and an ample opportunity to rethink the restrictions (or lack thereof) on these platforms.

Finally, transparency is vital in making the political decisions surrounding content moderation. Though recent blog posts from Facebook and Twitter have helped in this regard, I claim there are two goals to strive for in ensuring transparency in content moderation policies. The first goal is transparency in moderation decision making. Not only should the user know what content is causing them to be penalized, but also why the decision with their content was made. This is a big issue with AI moderation, beyond it being ineffective, if the AI is trained using a neural network, the actual decision making is effectively a black-box for all involved. This issue can primarily be resolved with a shift toward human-driven decision making. Having humans making content moderation decisions allows for us to ensure that any decision made within the system can be explained. The second goal is for there to be more transparency in the moderation policy making decisions. This is where private entities will have to play a larger role, as every platform is different, meaning they will need differing policies. As previously stated, Twitter and Facebook (among others) have used their blogs as a means of transparency, but they do not provide much information, and only give the broadest ideas of what any new initiatives are meant to do. Here, a public/private compromise can be struck, where legislation can be passed making these sites make both their moderation policies publicly available and any proposed changes to those policies publicly available. This would make it more clear as to what is being done, why it is being done, and how policy decisions are made. This will help public trust in the systems, as information on how moderation policies are made would be publicly available and consumable.

As we can see, both a shift in the role of government as well as the companies running social media sites are requisite to properly handling online content moderation. The ideas outlined here highlight broad goals which should be pushed for as legislation is drafted and sites change their moderation policies. Changes in the conception of free speech, changes to private moderation policies, and transparency on both the part of public and private institutions are all vital to ensuring this political issue is handled carefully and well. Broadening what constitutes unprotected speech by limiting protections for false information and hate speech will serve to protect civil discourse online. Additionally, changes to private content moderation means, both along those lines, as well as ensuring all decisions are made by human actors will ensure a more trustworthy system with fewer arbitrary decisions. Finally, requiring transparency in decision making will be vital to ensuring public trust in this decision making progress. This likely means a compromise struck between public and private institutions, with government holding sites accountable in keeping their policy decisions transparent and open to the public.

The ideal means of handling social media content moderation would be striking a balance between the responsibilities of government and the private corporations running those sites. A shift in the conception of free speech, at least online, which would render hate-speech and knowingly spreading misinformation unacceptable, in conjunction with a step up from social media sites in terms of both moderation and transparency would help to significantly improve the health of online discourse. I claim these policies will improve the health of online discourse, help social media sites maintain their audience, and strike the balance between public and private interests as the associated technology continues to develop.

Conclusion: The Content Moderation Compromise

This paper accomplishes two tasks. First, I claim that social media ought to be treated as a public good, following Emmanuel Mesthene's conception of technological development. Second, because content moderation is a public policy issue, I create a public policy framework which I claim will be optimal for handling content moderation on social media. The literature review conducted in the first two sections of this paper demonstrate the validity of the first claim, and give a basis for why the public policy proposal in the third section is important and valid.

The first literature review in this paper demonstrates the failures of current content moderation methods. Multiple case studies are examined to demonstrate the various missteps that have been taken by social media corporations. The second literature review demonstrates how as technology grows, its influence does too. As social media has exploded in popularity and use, it has gone from a private good affecting relatively few people, to a public good, affecting billions. Additionally, it means that political decisions regarding social media need to take into account the interests of both public and private stakeholders, as they both play a role in regulating this still developing technology.

This leads to the public policy framework. The fast growth of social media and its ubiquity have made producing effective content moderation policies vital to the continued operation of these websites. This is especially true as they continue to amass more users, and act as the go-to places for people to interact online. My policy ideas include a rethinking of American free speech policy, ensuring human actors make content moderation decisions, flagging false or misleading content, and transparency about the moderation systems themselves. This literature review and policy proposal provides a basic groundwork upon which a more specific and thorough content moderation policy can be laid, and serves to provide a template for which individual sites can work from and build upon. What I lay out is a basic framework from which specific public policy can develop, as well as places for further research in this field. Though some of the goals posed are broad and lofty, I believe they are worth striving toward, and as further research is done they can be further iterated upon.

References

Communications Decency Act of 1996, 47 U.S.C § 230 Protection for Private Blocking and Screening of Offensive Material (1996)

Dizikes, P. (2018, March 8). Study: On Twitter, false news travels faster than true stories. Retrieved October 09, 2020, from https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308

Facebook. (n.d.). *Working to Stop Misinformation and False News*. Retrieved March 4, 2021, from https://www.facebook.com/formedia/blog/working-to-stop-misinformation-and-false-news

Edelson, L., Nguyen, M., Goldstein, I., Lauringer, T., Goga, O., & McCoy, T. (2021, March 3). *Far-right news sources on Facebook more engaging - Cybersecurity for Democracy*. Medium. https://medium.com/cybersecurity-for-democracy/far-right-news-sources-on-facebook-more-engaging-e04a01efae90

Gey, S. (2008). The First Amendment and the Dissemination of Socially Worthless Untruths. *Florida State University Law Review*, *36*(1), 1–22. https://ir.law.fsu.edu/cgi/viewcontent.cgi?article=1128&context=lr

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, *7*(2), 205395172094323. https://doi.org/10.1177/2053951720943234

Kelion, B. L. (2020, November 17). *Facebook and Twitter grilled over US election actions*. BBC News. https://www.bbc.com/news/education-54974819

Knechtle, J. (2008). Holocaust Denial and the Concept of Dignity in the European Union. *Florida State University Law Review*, *36*(1), 42–63. https://ir.law.fsu.edu/cgi/viewcontent.cgi?article=1130&context=lr

Mesthene, E. G. (1970). *TECHNOLOGICAL CHANGE its impact on man and society* (2nd Printing ed.). Mentor / New American Library.

Moon, R. (2008). Hate Speech Regualtions in Canada. *Florida State University Law Review*, *36*(1), 79–95. https://ir.law.fsu.edu/cgi/viewcontent.cgi?article=1132&context=lr

Nguyen , C. T. (2019, October 31). The problem of living inside echo chambers. Retrieved October 09, 2020, from https://theconversation.com/the-problem-of-living-inside-echo-chambers-110486

Permanent suspension of @realDonaldTrump. (2021, January 8). Twitter. https://blog.twitter.com/en_us/topics/company/2020/suspension.html

Petkauskas, V. (2021, February 17). *70TB of Parler users' messages, videos, and posts leaked by security researchers*. CyberNews. https://cybernews.com/news/70tb-of-parler-users-messages-videos-and-posts-leaked-by-security-researchers/

Thrasher, S. W. (2018, December 7). *What Tumblr's porn ban really means*. The Atlantic. https://www.theatlantic.com/technology/archive/2018/12/tumblr-adult-content-porn/577471/

Zannettou, S. (2018, April). What is Gab: a bastion of free speech or an Alt-Right echo chamber. *WWW '18: Companion Proceedings of the The Web Conference*, 1007–1014. https://doi.org/10.1145/3184558.3191531