Christian Pak

Professor Lew

ENGL102 32675

5 May 2024

To Quell the Hate Machine

Advances in technology have radically revolutionized how humans live their daily lives. Communication, transportation, information, and entertainment are more efficient than ever and are still being improved. One of the most crucial developments in the modern age is the establishment of the Internet. This interconnected global network allows information to spread worldwide at unbelievable speeds. Additionally, social media platforms allow people to interact with millions of other users. However, concealed by this wondrous display of technological prowess lies a toxic underbelly. The anonymity of users and the lack of direct interaction allow the internet to be a dangerous catalyst for hate. As a result, numerous users spread malicious speech out into the world. Despite most platforms discouraging this malicious expression, hate speech still thrives on social media like a festering infection. In the end, online hate has prompted countries around the world to discuss how to deal with hate speech. While many people argue that hate speech is harmful and should be removed from the internet, many others believe censoring hate speech infringes on free expression. Additionally, defining hate speech has also become a prevalent topic among groups and individuals. Reaching conclusions to these debates is crucial for the sociopolitical environment facilitated by the internet. Ultimately, because online hate speech causes mental harm to others, leads to real-world violence, and is improperly handled by companies, governments should reach a consensus on the definition of

hate speech and create rules to prevent its proliferation despite fears of infringing on the freedom of speech.

First and foremost, concluding on the definition of hate speech is crucial to any legislation and jurisdiction around hate speech. Without a clear definition, ruling on hate speech cases becomes complicated and too open for interpretation. While many groups have created their definitions of hate speech, concluding on a single one is necessary. Additionally, while many definitions are similar, certain aspects must be addressed. One such definition was recommended by the Committee of Ministers of the Council of Europe. In “Social Media and the Kurdish Issue in Turkey: Hate Speech, Free Speech and Human Security" by Funda Gençoğlu Onbaşi, the author references their definition while discussing current interpretations of hate speech. She states that the council believes hate speech includes “all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance…” (117). In this instance, hate speech encompasses any expression that spreads hateful content involving certain groups such as race, religion, and nationality.

Similarly, the Human Rights Watch defines hate speech as “any form of expression regarded as offensive to racial, ethnic and religious groups and other discrete minorities, and to women" (Gençoğlu Onbaşi 117). This example, while comparable in the idea that it’s protecting the same groups, also specifies that any speech that is offensive to these groups is considered hate speech. This demonstrates how minute nuances in hate speech definitions alter how they can be interpreted. Another popular example of hate speech definitions can be found on social media platforms. In their rules, platforms such as Twitter, Facebook, and Instagram often discourage hateful conduct. In a summary of Twitter’s policy from 2021, hate speech is classified as “...hateful imagery and display names, violent threats, ‘wishing, hoping or calling for serious

harm,' as well as 'repeated slurs, tropes or other content that intends to dehumanize, degrade or reinforce negative or harmful stereotypes about a protected category'" (Hietan and Eddobo 449). Intent is implied to be a large factor in this definition. Whether or not someone engaging in hate speech intends to harm others with their expression is also something to consider when reaching a consensus on hate speech. In culmination, these examples highlight the variation in definitions and how they can drastically alter potential interpretations of hate speech. An absolute definition must be reached.

Subsequently, after overviewing current definitions, evaluating similarities and differences between them enables groups to narrow down definitions to key concepts, and select the best choice to agree on. Researchers Mika Hietan and Johan Eddobo, in their article "Towards a Definition of Hate Speech—With a Focus on Online Contexts," provide their suggestions. The pair concludes with four classifications after criticizing current definitions for being too broad and abstract. They label these categories as "a teleological, a pure consequentialist, a formal (in Platonic or Aristotelian terms), and a consensus or relativist mode" (451). Under these terms, Hietan and Eddobo explain specific ways hate speech can be defined. Teleological refers to the intent and the effects of the expression; pure consequentialist relates to the results of the expression alone; formal connects to the ideas expressed; consensus is built upon agreement on categorizing hate speech (451). These interpretations feature key differences between current definitions and separate them into clear specifications. Because most definitions agree on the victim groups of hate speech, little needs to be discussed in that regard. Instead, governments should focus on agreeing as to whether speech becomes hateful when it has malicious intent, when it results in hate-related consequences, or when it contains ideas deemed

as hateful. In doing so, organizations will then be able to create specific rules to identify and prevent hate speech from spreading on the internet.

Next, the main argument surrounding online hate speech involves the opposing ethical viewpoints of free speech and safer internet. Several uphold the principle of free expression as an absolute right that must never be infringed upon. Under this justification, these people defend hate speech as an expression protected by free speech. In addition, those against hate speech laws often fear that giving up any amount of speech will lead to organizations gaining more power over society. In "Chapter 4: Free Speech and The Internet" of the book *Threats to Civil Liberties: Speech*, this viewpoint is discussed. The author states that many fear more forms of speech will be construed to fall under hate speech and that social media companies could go against the public interest (Currie 20-21). These are common concerns held by free speech advocates. By taking control of what is considered hate speech, powerful groups could potentially manipulate more than just hate speech. Likewise, these sentiments are also shared by journalist H.K. Rivera who comments, "Once you cannot speak out without repercussions, all other rights are taken soon after. This is how dictators take over countries and oppress the citizens." Again, Rivera exemplifies the belief that free speech must not be violated. But while both texts offer opinions that capture the morality of defending freedom, they also fall under the slippery slope fallacy. In each case, the arguments rely on the idea that restricting hate speech will lead to more rights being taken away. However, the devastating consequences of hate speech outweigh the need to uphold freedom of expression.

One of these consequences is the harmful mental effects on a person's dignity and stress caused by malice-filled expression. As much as mankind should have the right to convey their thoughts and emotions at will, they also have the right to live without feelings of insecurity. Hate

speech almost always violates this natural right. Many researchers and scholars have introduced the topic of human security into discussion about hate speech. They define this term as "…the universal value of [the] equal right to live in dignity" (Gençoğlu Onbaşi 125). In other words, all humans naturally have the right to be respected by oneself and others. Because hate speech inherently breaks this principle, under the concept of human security, hate speech is morally wrong. In a similar vein, a study published under the title "Online Hate Speech Victimization: Consequences For Victims' Feelings of Insecurity" explores how digital hate speech affects people's sense of security. In the researchers' report, they point to pre-existing studies that found victimization and hate speech have a positive causal relationship with feelings of depression and loneliness (Dreißigacker et al. 2). Alone, this already serves as evidence for the negative effects of hate speech. Those exposed to it experience numerous unhealthy mental afflictions. However, within the study, Dreißigacker and other researchers investigate how online hate speech results in feelings of insecurity in comparison to offline hate speech. Their results conclude that online hate speech does lead to greater levels of insecurity among victims in contrast to offline hate speech (7). This highlights the urgency of reducing online hate speech. Because of the qualities that hate speech acquires when communicated on the internet, the effect leads to even more harm that must be addressed. These clear detrimental effects that digital hate speech has on people's self-esteem and dignity explicitly show the importance of removing such malicious expressions.

Another example of hateful expression wounding the mental well-being of victims is the stress it causes. Feelings of stress and anxiety have been proven to increase through exposure to hate speech. In a recent study, three Georgia Tech researchers collected data from the social media platform Reddit to determine how hate speech affects stress expression among college students. In their report "Prevalence and Psychological Effects of Hateful Speech in Online

College Communities," they record that "the stress level of the Treatment users (mean=139%) is higher than the Control users (mean=106%)... [suggesting] that this exposure likely has a causal relationship with the online stress expression of the users" (Saha et al. 14). In this instance, "treatment users" refer to those exposed to hate speech, while "control users" refer to those not exposed. According to the data presented, there is a clear increase in stress among college students who experience hateful expressions. Because unnecessary stress can lead to more harmful effects on both the mind and body, the data collected in this experiment demonstrates the negative outcomes of hate speech. Overall, this example portrays another way in which digital hate speech proves detrimental to the mental health of its victims, and why the government must work to remove it.

Finally, the last reason malicious speech needs to be removed from the internet is that its spread can incite physical hate crimes. Although this topic has yet to be widely explored, research has indicated a connection between hate speech and hate crime. In "Hate Speech on Twitter Predicts Frequency of Real-Life Hate Crimes," the writer discusses a study conducted by NYU researchers that links online hate to violence in the real world. The writer quotes the lead researcher, Rumi Chunara, who stated that "'...they found that more targeted, discriminatory tweets posted in a city related to a higher number of hate crimes.'" This positive relationship implies a close correlation between the two forms of hate. Therefore, by reducing digital hate speech, the world will also likely see a decrease in hate-driven crimes. In all, this relationship serves as more evidence to support the importance of censoring hate speech.

After demonstrating why hate speech should be censored, the next aspect of the issue is who should be responsible for the role of removing it. Many individuals find the idea of granting the government the power to censor information unappealing. As discussed previously many

worry that granting the government this power will allow them to control the spread of information. Those who agree with this sentiment may instead look to social media companies such as Twitter and Facebook as the lesser of two evils. Some instances do demonstrate successful censorship by social media platforms. For example, in 2017, after a rally in Charlottesville, Virginia, turned violent, various social media platforms exploded with violent threats and insults toward Jews, Muslims, and African Americans. In response, over the next, these platforms began banning accounts that were engaging in this hate expression (Currie 20). The response by social media companies in this instance portrays successful censorship. Their response was swift, and they established their position against the spread of hate speech. This indicates that platforms have the power and capability to respond to hate speech.

However, on numerous other occasions, these same companies have proven to be unreliable in their efforts to remove hate speech. Furthermore, they instead seem more interested in their benefit over the well-being of their networks and users. In the articles titled “How Facebook Hides How Terrible It Is with Hate Speech” by Noah Giansiracusa and "Facebook Is Even Worse than Anyone Imagined" by Alex Shepard, the two reveal underhanded practices committed by Facebook in handling hate speech on its platform. Giansiracusa exclaims that while Facebook claims “...that it removes more than 90 percent of hate speech on its platform… in private internal communications the company says the figure is only an atrocious 3 to 5 percent.” This not only reveals the ineffectiveness of Facebook’s censorship but also demonstrates how social media platforms care more about their image than correcting the underlying issue. Giansiracusa then reveals that the data they presented is technically true, but measures a different ratio. Because they hid the relevant data with this irrelevant one, it further demonstrates the lengths Facebook will go through to manipulate information to improve its

image. Likewise, Shepard points to a leak revealing that executives at Facebook are aware of the malicious content on the platform and the harm it causes but refuse to reduce it, instead focusing on profit. Once more, Facebook represents how greed makes corporations untrustworthy, and thus not reliable enough to censor hate speech. To resolve issues involving hate expression, the law must instead be used to enforce rules against hate speech or hold companies accountable for the proper removal of hate.

Lastly, one final argument to contrast the idea that hate speech should be censored is because the nature of the internet prevents proper jurisdiction. Due to certain intrinsic qualities of the internet and social media, some believe the government won't be able to enforce laws. They are concerned that many users will find ways to bypass rules and continue to spread hate speech. A researcher at the Oxford Internet Institute, Bharath Ganesh, conveys this concept through his journal article "The Ungovernability of Digital Hate Culture." He attributes the ungovernability of the internet to the three properties of "...its decentralized structure, its ability to quickly navigate and migrate across websites, and its use of coded language to flout law and regulation" (36-37). To explain, because social media comprises several platforms that all use different rules, shifting between sites is an effective way to avoid regulation and is guaranteed to be exploited. Then, by inventing new terms on the internet such as "white genocide" and "rapefugees", users can further avoid punishment while spreading the same themes of hate (Ganesh 39). While these points have merit, they also highlight the importance of reaching a consensus on the definition of hate speech. Reaching a single specific interpretation of hate speech would allow a common definition to apply to the entire internet, thus limiting the ability of malicious users to migrate between sites. Additionally, by settling on a definition that can adapt to the fast-changing internet, new terms can quickly be labeled as hate speech as well. Lastly, even if hate speech

can’t be removed from all corners of the internet, simply reducing its proliferation on popular platforms will greatly reduce the harm it inflicts on the world. In summary, while the internet provides some unique challenges for creating hate speech laws, reaching a consensus on the definition of hate speech will effectively resolve the majority of them.

To conclude, digital hate speech is a form of malicious expression that plagues social media platforms. With its rise to prominence, many have been discussing how hate should be handled. While many wish to protect hate speech under the freedom of speech or want to leave censorship in the hands of private companies like Instagram, Twitter, and Facebook, due to detrimental mental effects and real-world violence that result from hate speech, in addition to the unreliability of social media platforms, governments must agree on a definition of hate speech and create laws to mitigate the spread of hate on the internet. The current variation and vagueness in hate speech definition greatly alter how different groups identify it. Thus, establishing a widely accepted definition is necessary for further discussion into solving the hate speech issue. Then because hate speech results in increased stress, insecurity, loneliness, depressive emotions, and real-world violence, governments should focus on enforcing hate speech laws. Lastly, because social media companies engage in practices that benefit themselves rather than resolve hate speech issues on their platforms, they can not be trusted with the role of hate speech censorship, meaning the government has to create hate speech laws instead. As the world develops further into a digital society, global policy must adapt to the internet. Due to internet culture reshaping millions of human lives, many old laws and legislation struggle to be enforced in the digital landscape. Hate speech is only one issue being discussed. By resolving each one, the internet can gradually become a safer place for today's and tomorrow's users.

Works Cited

Currie, Stephen. "Chapter 4: Free Speech and The Internet." *Threats to Civil Liberties: Speech.* ReferencePoint Press, Inc, 2020, pp. 19-22. *ProQuest*, https://montgomerycollege.idm.oclc.org/login?url=https://www.proquest.com/books/threats-civil-liberties-speech/docview/2421609409/se-2. Accessed 13 April 2024.

Dreißigacker, Arne, et al. "Online Hate Speech Victimization: Consequences For Victims' Feelings of Insecurity." *Crime Science*, vol. 13, no. 1, Dec. 2024, p. NA. *Gale Academic OneFile*, dx.doi.org.montgomerycollege.idm.oclc.org/10.1186/s40163-024-00204-y. Accessed 4 May 2024.

Ganesh, Bharath. “The Ungovernability of Digital Hate Culture.” *Journal of International Affairs*, vol. 71, no. 2, 2018, pp. 36–41. *JSTOR*, https://www.jstor.org/stable/26552328. Accessed 20 Apr. 2024.

Gençoğlu Onbaşi, Funda. “Social Media and the Kurdish Issue in Turkey: Hate Speech, Free Speech and Human Security.” *Turkish Studies*, vol. 16, no. 1, Mar. 2015, pp. 115–30. *EBSCOhost*, https://doi-org.montgomerycollege.idm.oclc.org/10.1080/14683849.2015.1021248.

Giansiracusa, Noah. “How Facebook Hides How Terrible It Is with Hate Speech.” *Wired*, Conde Nast, 15 Oct. 2021, www.wired.com/story/facebooks-deceptive-math-when-it-comes-to-hate-speech/. Accessed 9 April 2024.

“Hate Speech on Twitter Predicts Frequency of Real-Life Hate Crimes.” *NYU*, 24 June 2019, www.nyu.edu/about/news-publications/news/2019/june/hate-speech-on-twitter-predicts-frequency-of-real-life-hate-crim.html. Accessed 21 April 2024.

Hietanen, Mika and Eddebo, Johan. “Towards a Definition of Hate Speech—With a Focus on Online Contexts.” *Journal of Communication Inquiry*, vol. 47, no. 4, 2022, pp. 440-458. https://doi-org.montgomerycollege.idm.oclc.org/10.1177/01968599221124309. Accessed 21 April 2024.

Rivera, H.K. "Is 'Hate Speech' Dangerous?" *Gale Opposing Viewpoints Online Collection*, Gale, 2024. *Gale In Context: Opposing Viewpoints*, link.gale.com/apps/doc/EZFJMV224534984/OVIC?u=rock77357&sid=bookmark-OVIC&xid=55decd4d. Accessed 11 Apr. 2024. Originally published as "Is 'Hate Speech' Dangerous?" *American Thinker*, 19 Oct. 2019.

Saha, Koustuv, et al. “Prevalence and Psychological Effects of Hateful Speech in Online College Communities.” *Proceedings of the ... ACM Web Science Conference. ACM Web Science Conference* vol. 2019 (2019): 255-264. doi:10.1145/3292522.3326032. Accessed 20 Apr. 2024.

Shephard, Alex. "Facebook Is Even Worse than Anyone Imagined." *Gale Opposing Viewpoints Online Collection*, Gale, 2024. *Gale In Context: Opposing Viewpoints*, link.gale.com/apps/doc/EARLYV381805389/OVIC?u=rock77357&sid=bookmark-OVIC&xid=f00331c6. Accessed 26 Apr. 2024. Originally published as "Facebook Is Even Worse Than Anyone Imagined," *The New Republic*, 25 Oct. 2021.