Content Moderation as an Evolving System:

A Comparison of Societal Values of Free Speech and Public Safety

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

Madeleine Ashby

Fall 2022

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 Assignments

Advisor

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

> There is held to be an asymmetry whereby the societal costs of discussing certain topics inevitably outweigh any benefits from doing so. No such asymmetry has been empirically demonstrated, and stifling debate around taboo topics can itself do active harm. (Carl, 2018, p399)

Introduction

Rooted in implicit bias, hate speech surges through society, primarily in the form of social media posts. The most immediate effects of this are obvious – increased racial tensions and a magnified sense of isolation, ultimately jeopardizing the cohesion of society. Unopposed, online hate speech can facilitate the establishment of echo chambers which lead to further polarization and mob mentality. In more extreme cases, communities mobilize and the hate shifts offline, to real life, where people are driven to commit violent acts. Lavin (2018) argues that the existence of unregulated platforms such as 4chan and Gab serve to amplify these problems, and the murders of several members of minority groups on behalf of neo-Nazi hub The Daily Stormer only further exemplify this point.

As these hate crimes continue to increase in frequency, platforms all over the internet are taking a stance. Sites such as Instagram and Facebook have implemented content moderation algorithms that remove posts which violate community guidelines. Facebook (n.d.) claims that their standards for moderation are designed to "create a safe environment." Other sites, including Gab and 4chan, refuse to follow suit, arguing that to do so would be a violation of freedom of speech. Gab's creator, Andrew Torba (2021), posted, "we believe that a moderation policy which adheres to the First Amendment, thereby permitting offensive content to rise to the surface, is a valuable and necessary utility to society." At the intersection of such polarization regarding content moderation lie two groups of people: those against content moderation, and

those for it. The caveat here is that the moderation that both Facebook and Torba reference is constructed under the assumption that we can easily distinguish between what is hate speech and what is not. As this is not the case, it is important to account for both the perspectives for and against content moderation. In doing so, it becomes clear that the real challenge is to find an approach to the debate which balances the concerns of each group. In this paper, I utilize Pinch and Bijker's Social Construction of Technology framework and ideas from Noah Carl's "How Stifling Debate Around Race, Genes and IQ Can Do Harm," to evaluate these two opposing arguments around content moderation and determine which value – public safety or freedom of speech – should be prioritized in order to reform content moderation to be more than a flawed "fix-all" solution.

Problem Definition

The Propagation of Hate Speech and the Implications of its Moderation

The unconscious mind, although entirely composed of naturally occurring processes unavailable to introspection, plays a remarkable role in affecting behavior and emotions (Ricee, 2021). As such, the brain is capable of making very impulsive decisions without conscious awareness or understanding. The brain learns from what it is presented with, even encoding information into the unconscious that has not yet passed logical validation in conscious thought. Consequently, your mind makes subconscious, illogical, yet quick determinations based on whatever information you are exposed to, regardless of that information's veracity. This can result in impulsive thoughts and behaviors as reactions to those stimuli. For instance, if you are raised in a neighborhood in which all families are of the same culture, you would likely be more receptive to familiar-sounding names than to those of other cultures ("Implicit Bias", 2020). This predisposition which leads to a cognitive "shortcut" is called implicit bias.

One of the most prominent manifestations of implicit bias is hate speech: where groups of people are discriminated against and targeted with vicious stereotypes and violent ideas. As such, there are many types of negative health effects associated with hate speech. Cramer et al. (2020, n.p.) list effects such as: poor emotional well-being; feelings of shame and fear; poor mental health, including depression, anxiety, and posttraumatic stress; and poor physical health, including physical injury and stress. The Anti-Defamation League's annual survey suggests the following additional effects shown in Figure 1:

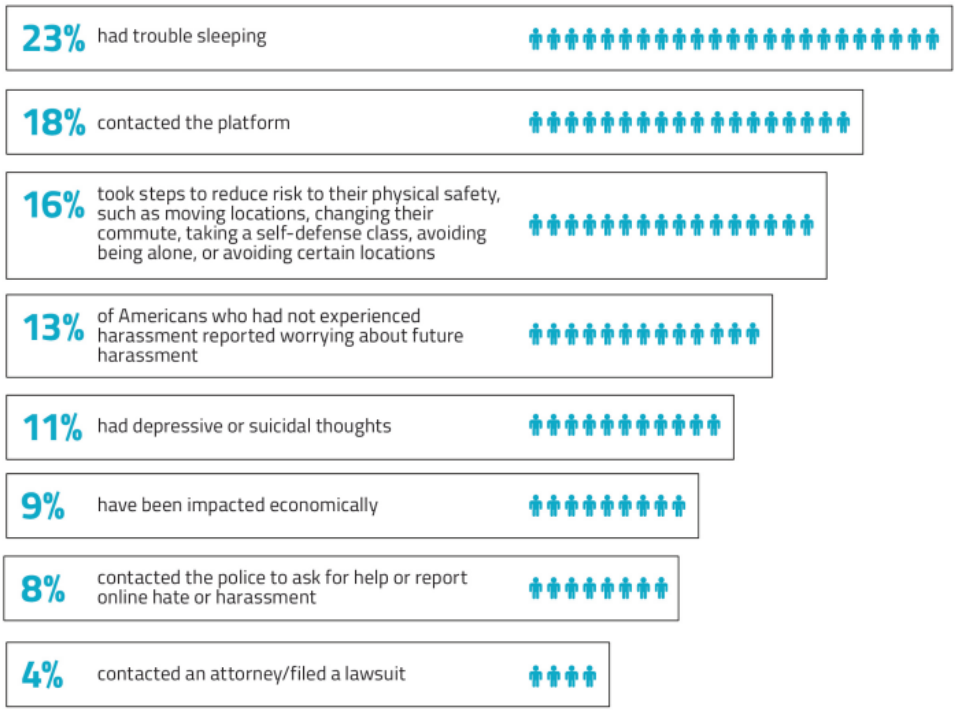

Figure 1: Impacts of Online Hate Speech

Those affected by hate and harassment on social media reported unexpectedly high percentages of difficulty sleeping and depressive thoughts based on the Annual Survey of Hate and Harassment on social media (Anti-Defamation League, 2021)

Furthermore, since this hate is often targeted at marginalized groups, a single hateful act can victimize an entire community. While it is understandably difficult (and potentially dangerous) to stop hate speech when it occurs in face-to-face scenarios, it is becoming increasingly important to put an end to it before it becomes violent. A good place to start is on social media.

The existence of hate speech on social media is no surprise – it's an explicit projection of prejudice, first conceived when those with implicit biases are radicalized, then resulting in the development of implicit bias in others and eventually, radicalization. In other words, the propagation of implicit biases into hate speech (and vice versa) is purely cyclical. Hate speech flows through social media, and although some social media sites implement algorithms to flag and limit such hate, these imperfect algorithms feed back into the propagation of implicit bias. Consequently, hate speech is the pinnacle of a growing cycle of implicit bias that augments hate and discrimination, all the while contaminating society's view of reality with growing personal biases, only leading to more hate. Implicit bias is the root of hate, and efforts to stop hate speech are being thwarted by the biases of those fighting against it.

The twentieth and twenty-first centuries have seen the increase in the mobilization of radical hate groups and the hate crimes they have committed. One example is the mass shooting at the Tree of Life Synagogue in 2018. Researchers at the Anti-Defamation League have argued that the tragedy should not have been surprising; perpetrator Robert Bowers had been posting dehumanizing and threatening anti-Semitic messages on social media site Gab for weeks leading up to the event (Guthrie, 2018). Several similar acts have occurred since, including the attack at the ChristChurch Mosque in New Zealand in 2019. According to Dr. Adriel Koch (2020), the high accessibility of platforms like Gab is the driving force behind widespread mobilization of extremist groups.

In light of these events, progress towards comfortable, hate-free platforms has been made through the creation of hate speech detection algorithms, but this task hasn't been easy. According to the latest results from the Anti-Defamation League's (2021) annual survey of hate and harassment on social media, "only 14% of those who experienced a physical threat [on social

media] said the platform deleted the threatening content," and "just 17% of those who experienced a physical threat to the platform stated that the platform blocked the perpetrator who posted the content." Both manual and automated hate speech moderation are being implemented by many social media companies; posts reported by users are aggregated and sorted into queues for the moderators to review, and a quick binary decision is made: the post is offensive and hateful, or it isn't (Chotiner, 2019). Naturally, this moderation poses a few substantial issues. Firstly, the moderation is biased. Widespread disagreements regarding the definition of hate speech serve as an obstacle to creating an algorithm that detects it (MacAvaney et al., 2019). This results in biased data that makes it difficult to train a model that predicts and acts fairly. Moreover, automated moderation utilizes algorithms that often neglect context and are consequently highly sensitive models that misclassify both hateful and innocuous posts (Dawson, 2020). Secondly, the moderators are biased. Manual moderation makes use of humans who are inherently biased and become desensitized to the content over time (Chotiner, 2019). Thirdly, the process of removing a user's thoughts and opinions from a platform is sometimes perceived as an infringement upon freedom of speech, possibly even further radicalizing those affected by censorship and their followers. A summary of these problems are outlined in Figure 2.

Problem	Cause
Moderation is biased.	Disagreement surrounding what constitutes hate speech makes it difficult to accurately detect it.
	Automated algorithms neglect the context in which the supposed hate speech is used.
Moderators are biased.	Humans are subject to implicit bias which can be subconsciously exercised when deciding if a post is harmful or not.
Disagreement surrounding classification and removal of harmful or harmless speech can lead to further radicalization.	Removing a user's opinion can be seen as infringement upon freedom of speech.

Figure 2: Current Problems with Content Moderation and their Potential Causes
Table describing the problems currently associated with content moderation and their apparent problems based on a broad overview of available literature. (Created by Author)

This study aims to analyze the current state of content moderation and its implications on social media users in order to build a deeper understanding of how different societal values have shaped its evolution and in which situations certain values might need prioritization over others. This is achieved through an ethical analysis comparing two opposing arguments: the first argument is that the reinforcement of extremist ideals on online platforms is a threat to public safety and should be moderated, while the second argument is that the moderation of hate speech is imposing limits on freedom of speech, thus silencing dissent in marginalized groups.

Methods

At this point, it is clear that some degree of moderation is necessary in order to safely diminish hate on social media platforms, and this conversation is one that will continue to elicit emotional responses on both sides; however, this debate is essential when it comes to finding the Golden Mean – a notion conceived by Aristotle that constitutes the balance between the vices of excess and deficiency. In this case, excess is found in over-moderation – opposed by those valuing free speech – and deficiency is found in under-moderation – opposed by those valuing

public safety. To find the Golden Mean is to find the balance between the two arguments which will facilitate progress in a direction that considers content moderation as a central component of a larger system rather than on its own.

As previously stated, this ethical analysis utilizes Pinch and Bijker's Social Construction of Technology (SCOT) framework and draws on ideas presented by Noah Carl in his theoretical article, "How Stifling Debate Around Race, Genes and IQ Can Do Harm."

The Social Construction of Technology framework is outlined in Pinch and Bijker's paper, "The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other" (1984). This framework discusses technological development as a system, maintaining that technology shapes and is shaped by society. This means that technology is never a standalone device, but instead that all technology lies in a larger system with several components. This concept is illustrated in the figure below. The central component is technology, and the surrounding components are the social groups which technology affects, the perceived problems of each social group, and the potential solutions to each problem. When all of the components are connected, they form a complex and multi-directional web which represent the SCOT system.

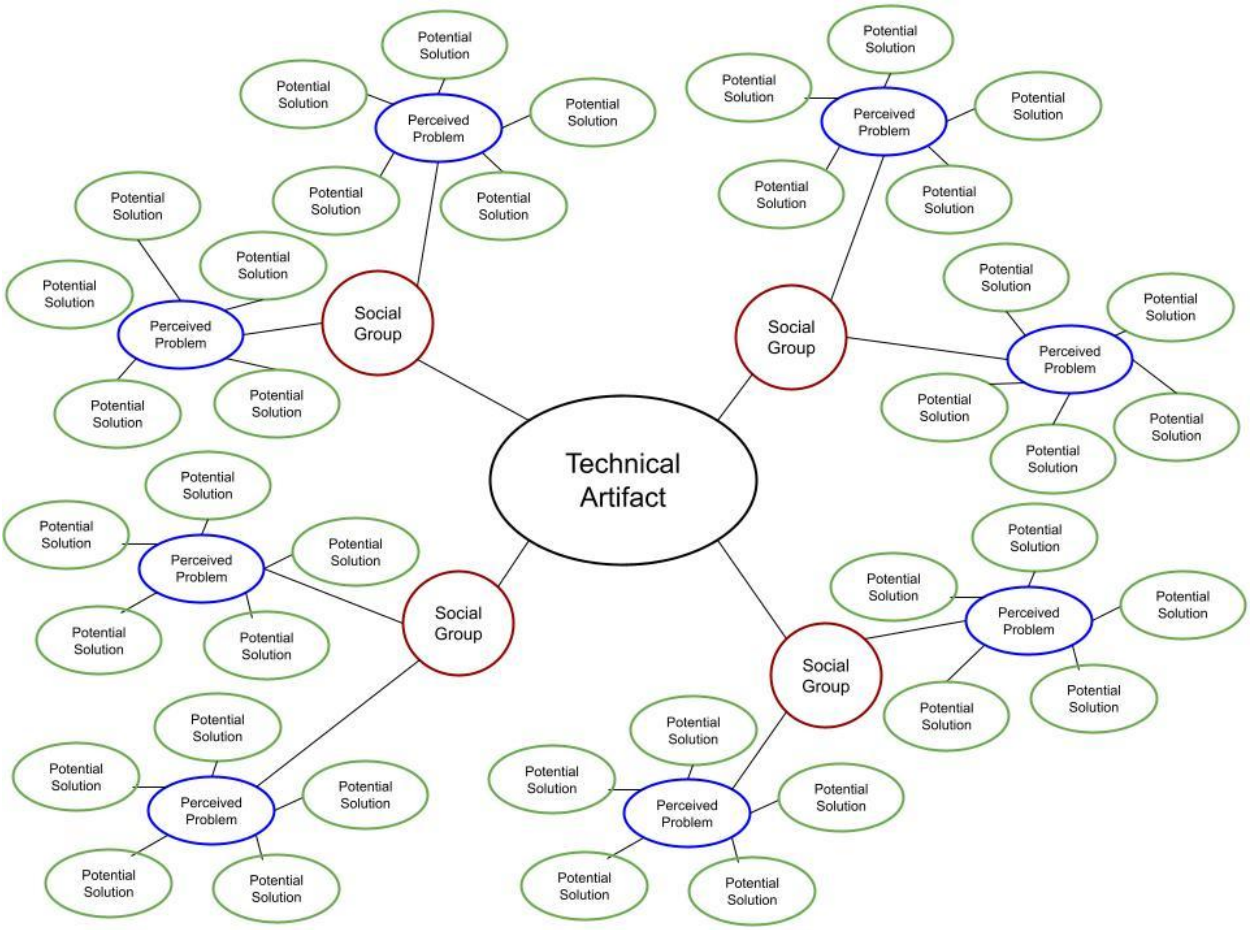

Figure 3: Technology as a System

The Relationship between Technology and Relevant Social Groups, between Relevant Social Groups and their Perceived Problems, and between the Perceived Problems and their Potential Solutions (Created by Author)

Overall, SCOT is a highly applicable framework for analyzing technological systems because it emphasizes the distinction between the existence of a technology and the use of a technology. As such, there is a dependency on acknowledging this difference in order to develop a comprehensive understanding. In order to understand the use of a technological device, one must first understand the context in which it exists (Klett, 2018). In other words, it is critical that we understand the system of content moderation as a separate entity from the content moderation algorithms themselves so that we can properly assess their usefulness.

Noah Carl is a British sociologist and intelligence researcher who has published several papers exploring taboo and controversial social issues. In his article, "How Stifling Debate Around Race, Genes and IQ Can Do Harm," Carl discusses the perceived asymmetry between costs and benefits of debating taboo topics and asserts that such asymmetry does not exist. First, he claims that equating scientific statements with political discrimination holds our morals hostage to facts. Conversely, this means that the act of associating facts with political narratives results in the belief that a statement cannot be true if it has unpleasant moral implications. Second, he claims that the "blank slate" view of human nature has negative implications, where "blank slate" refers to the idea that all humans, presented with the same moral compass and resources, will develop the same ethical map – insinuating that humanity can be "perfected through appropriately targeted state intervention" (Carl, 2018, p403). Finally, he asserts that stifling debate around taboo topics has done immense harm to individuals and societal institutions by causing authorities to sweep problems under the rug to avoid uncomfortable conversations.

This paper utilizes the SCOT framework and Carl's assertions by analyzing two prominent social groups, their perceived problems and potential solutions, and comparing each group's arguments to Carl's arguments. This comparison serves to strengthen understanding of similarities in both arguments which may provide a foundation for compromise and movement towards an all-inclusive system. This method is shown in Figure 4.

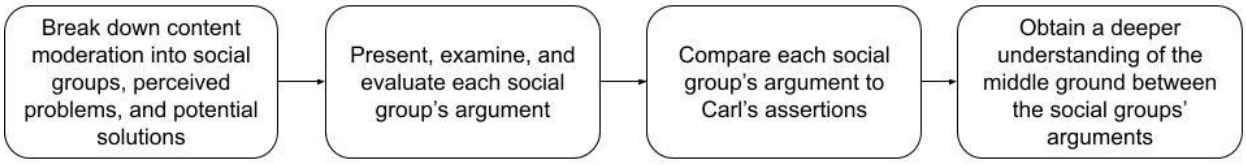

Figure 4: Method Map
A combination of SCOT framework and comparison to Carl's assertions (Created by Author)

An Ethical Analysis of Content Moderation

In Support of Stricter Content Moderation

The first component of the system I analyze is the argument presented by those in favor of stronger content moderation. This argument maintains that the reinforcement of extremist ideals on online platforms is a threat to public safety and should be eliminated. In order to evaluate this position, I examine two examples of threats to public safety and how content moderation (or a lack thereof) affected the outcome in each situation: the Unite the Right rally and Elsagate.

The Unite the Right rally in downtown Charlottesville, Virginia took place on August 11th and 12th, 2017, during which hundreds of white nationalists rioting over plans to remove a Confederate statue were met by counter-protesters. On the evening of August 11th, white nationalists supporting the alt-right movement marched through the University of Virginia's campus carrying burning torches with chants of "blood and soil" and "Jews will not replace us" (Elliott, 2022). In response, students and other counter-protesters gathered at the base of the Thomas Jefferson statue in front of the Rotunda with a banner that stated "VA Students Act Against White Supremacy" (Katz, 2017). The nationalists surrounded the students with their torches. On the morning of August 12th, swarms of people – both nationalists and counter-protesters – gathered around the Confederate General Robert E. Lee statue in the park downtown. The rally quickly turned violent when Neo-Nazi James Fields plowed his car into counter-protester pedestrians in the streets surrounding the park, murdering 32 year old Heather Heyer.

Talia Lavin (2018) argues that this rally and the tragedy that ensued was rooted in hate speech found on neo-Nazi hub *The Daily Stormer*. In July 2013, Andrew Anglin created *The*

Daily Stormer – named after the Nazi propaganda sheet, *Der Stürmer* – to be an American platform on which to spread radical white supremacist ideas. At its core, the website was a message board which actively invited its users to contribute to its mission according to a very strict style guide written by Anglin himself. According to this style guide, the ultimate goal was to "spread the message of nationalism and anti-Semitism to the masses" (Feinburg, 2017). This was done in a way that deliberately confused readers by juxtaposing a light, sarcastic tone with vitriolic, hateful material. In addition to the website itself not having any content moderation policies, the website's host, Cloudflare, also did not wish to police its clients' content (Nichols, 2017). However, this neutral position did not last long, for the website saw its permanent removal on August 16th, 2017, just four days after the tragedy in Charlottesville (Prince, 2017).

The Daily Stormer is a prime example of the radicalization of hate speech into real-life violence. Anglin's style guide highlights the nature of his website and the white nationalist users very well; they deliberately speak such outrageous ideas into existence that the readers doubt their sincerity. They want the readers to question the statements, to wonder if they are to be taken seriously. In turn, the readers' subconsciouses are manipulated into developing implicit biases and following the path toward white supremacy. As such, it is clear why pro-moderation groups would argue that this content should be moderated in order to prevent the growth and development of extremist groups.

The Elsagate scandal is another stark example of content moderation being a necessary technology in order to preserve public safety. Elsagate is a term coined by Reddit users in 2017 after the mass discovery of disturbing videos on YouTube. Using children's favorite characters, inappropriate content is targeted at children. Some of the most commonly presented themes include drug and alcohol abuse, cannibalism, gore, rape, and murder (Copia Institute, 2021). The

Elsagate phenomenon is particularly interesting because it managed to slip through the cracks of YouTube's content moderation algorithms. YouTube offers a version which features parental controls, pre-approved content, and intense video filtering to eradicate objectionable content (Townsend, 2018). Elsagate avoided this filtering by including innocuous keywords such as "education" and "learn." These disturbing videos should never have made it to YouTube Kids given the platform's high censorship. At a minimum, this incident calls for more transparency in YouTube's content moderation policies and algorithms in addition to an increase in human moderation as opposed to reliance on automated moderation.

After investigation of these two phenomena, it has become increasingly clear that there is validity to the pro-moderation argument. In the Unite the Right case, the implementation of content moderation by *The Daily Stormer* or Cloudflare could have flagged and removed posts that incited such violence; at the very least, this would have slowed the spread of the hate and possibly gotten legal attention that could have saved lives. In the Elsagate case, more transparency and human involvement in content moderation processes could have prevented such disturbing videos from surfacing on the internet. In the context of SCOT, the pro-moderation group has perceived problems of hate speech radicalizing into hate crimes, and a proposed solution to this is content moderation. However, this is only one possible branch of the entire web of content moderation. In order to obtain a more holistic understanding of content moderation as a system, we must evaluate the anti-moderation argument next.

Against Heavier Content Moderation

Those against content moderation form another social group; in understanding this perspective, including its problems and solutions, we are able to better grasp the debate around content moderation as a whole. This argument maintains that the moderation of hate speech on

social media platforms is imposing limits on freedom of speech, thus silencing dissent in marginalized communities. In order to evaluate this position, I examine two examples of moderation by which communities of people felt marginalized: the Al-Aqsa hashtag, and Twitter suspending Mariam Barghouti's account.

One of the holiest sites in the Islamic faith, the Al-Aqsa Mosque is a sanctuary to many Palestinan worshippers. In May 2021, the Mosque was repeatedly raided by Israeli police forces, during which users worldwide posted about the attacks on Instagram and Facebook using the hashtag "#AlAqsa" to raise awareness and share their devastation. However, Instagram flagged and removed many of these posts through its automated content filtering algorithm because it incorrectly identified the hashtag as referencing terrorism (Eidelman et al., 2021). When Instagram employees learned of the content removals and the company's justification for them, many filed grievances. In one case, a post was taken down because Facebook had classified "#AlAqsa" as a type of "dangerous individual and organization" (Mac, 2021). Although some content was eventually restored following these complaints, the fact of the matter still stands: "Palestinians and their supporters were silenced by one of the most powerful communications platforms in human history at a critical moment" (Eidelman et al., 2021).

In a similar turn of events, during the conflicts between Israeli forces and Palestinians in May 2021, writer and researcher Mariam Barghouti reported her experiences from the West Bank on social media. She posted several Tweets recounting her experiences and the events she witnessed until she posted "I feel like I'm in a war zone," resulting in her account being restricted due to misclassification in the moderation algorithm (Copia Institute, 2021).

These events are significant because they reiterate the need for more manual moderation in larger corporations. Users against heavier moderation have found these blanket decisions

made by automated algorithms to be curtailing speech in that the rules regulating such content are often highly subjective and consequently open the door to biased enforcement. The question these events raise is whether or not this is intentional – are these companies silencing dissent in order to pave the narrative they wish to create?

Results

As previously stated, Carl asserts that there are three fallacies that generally apply to debates around controversial topics. These assertions also apply to content moderation and the debate around it – in fact, three primary insights arise. The first insight that emerges from this analysis is that Carl's assertions regarding holding our morals hostage to the facts and having a "blank slate" view of humanity apply to the argument in support of stronger moderation. The second insight is that Carl's assertion that stifling debate around taboo topics is harmful is highly applicable to the argument against heavier moderation. The third insight is that Carl's fallacies are relevant to the debate about content moderation as a whole.

Comparing Carl's Ideas with the Pro-Moderation Argument

The pro-moderation argument moderately aligns with Carl's assertions in that the general concepts discussed by both are similar, but each's interpretations are different. For instance, content moderation can be thought of as a method of creating the "blank slate" that Carl defines to be a fresh start, untouched by human influence and experience. When content is removed, the platform is "clean" of that influence. Those in favor of content moderation argue that this blank slate is beneficial, that a platform which has been cleared is better than one which has been colored with hate. Moreover, they believe that a fresh start will eradicate any possibility of online hate resurfacing, for how can hate evolve when its roots no longer exist? However, Carl argues that the "blank slate" is not a plausible solution to the propagation of hate, but instead is

an extreme misuse of power to "remake humanity" (Carl, 2018). To wipe an event or collection of events from the record is to destroy our ability to construct our own future. With no way to reflect on the past, there is no way to learn from mistakes and consider societal values when creating new technologies.

In response to the viewpoint that a lack of moderation results in racism and hate crimes, Carl argues that to equate technological advancement (or lack thereof) to political or social agendas is to hold our morals hostage to facts. He describes the distinction between algorithmic decisions and normative conclusions in that algorithmic decisions are not perfectly indicative of society's morals and beliefs. This understanding is imperative for building a content moderation system that benefits society in the future. Furthermore, Carl explains that to worry about a hateful online community simply because you believe content moderation is not strong enough is a self-fulfilling prophecy which implicitly encourages hate. In addition, to say that content moderation is responsible for the existence of hate speech on social media platforms is naïve; it is the users who create hateful posts and should be held accountable for their actions. Consequently, in its current state, content moderation seems to be more of a technological fix than an actual solution.

The term "technological fix" was coined by Alvin Weinburg in 1965 to refer to a piece of technology that is created for circumventing social, political, or cultural problems (Johnston, 2018). In most cases, this has come to mean that the technology is a perceived solution to a larger societal problem, but that the technological solution does not actually solve the problem in its entirety. This is often due to the engineers of the new technology neglecting to consider the larger system in which it will lie, thus disregarding the social groups and values that have shaped and are shaped by it. Based on my findings thus far, it cannot be that the development of content

moderation is a "fix all" solution which marks an inflection point in hate speech history. Instead, content moderation earns its technological fix title due to its seeming lack of consideration for ethical responsibility.

Comparing Carl's Ideas with the Anti-Moderation Argument

The anti-moderation argument aligns with Carl's assertions to a greater extent in that both claim that silencing and censoring people causes active harm to both the silenced communities and the rest of society. To illustrate his point, Carl gives the example of British law enforcement failing to intervene and stop young girls from being groomed due to fear of being called "racist":

> [This evidence] illustrates that throwing around unsubstantiated charges of 'racism' can create a climate of fear in which people feel too paralysed to act, and that insofar as this is the case, doing so should not be considered a sensible precaution, … but potentially an unethical thing to do. (Carl, 2018, p405)

The phenomenon which Carl describes in this passage is political correctness, which is the careful avoidance of forms of expression or action that are perceived to marginalize discriminated groups. Nevertheless, the societal construct of political correctness does not constitute a moral compass and can still result in unethical acts in an effort to remain politically correct. This logic can be applied to content moderation as well. When morally correct social media users are censored by content moderation in an effort by the platform to appear politically correct, these users are silenced and the companies are not acting in alignment with ethical good. Once again, being politically correct does not constitute being morally correct.

Facebook has also fallen victim to this political correctness predicament. For years, they have consistently removed posts referencing Palestinian and Israeli conflict, inspiring uncertainty as to whether their content moderation is actually based on preservation of public safety, or if

other factors influence the removal of posts and users. At the very least, content moderation is inconsistent and confusing. Human-driven content moderation relies heavily on automated algorithms that neglect context and community policing (users reporting other users for perceived violations of guidelines). Consequently, some users are bound to be impacted more heavily than others. For example, users with larger accounts or accounts that post often are more likely to be reported than inactive users. Additionally, since users are policing each other, inconsistencies in reporting are inevitable; a post that offends one user may be humorous to another. This lack of consistency instills doubt towards the efficacy of content moderation.

Understanding Carl's Ideas in the Context of Content Moderation as a System

Figure 5 reiterates how each of Carl's assertions are adopted and mirrored by each group in regards to content moderation.

Carl's Assertion	**Application to Pro-Moderation Argument**	**Application to Anti-Moderation Argument**
Associating facts with political narratives results in facts being considered false if they have unpleasant moral implications.	By failing to moderate online hate, companies are supporting the hate crimes that ensue.	
A "blank slate" view of human nature has negative implications and can result in oppression.	A "blank slate" can help to ameliorate hate: a platform without hate is better than one covered in it.	
Stifling debate around taboo topics does active harm.		Silencing dissent is unethical and goes against fundamental rights.

Figure 5: Summary of Results

An explanation of each group's interpreted application of Carl's assertions in their own arguments (Created by Author)

It is equally as important to understand that Carl's claims are highly applicable to the debate around content moderation as a whole. There is no asymmetry between the costs and benefits of discussing content moderation; as Carl asserts, stifling debates on controversial topics can do

active harm. Moreover, debate around content moderation on social media is not unique in this context and even reflects many of the fallacies Carl describes about debate. To discourage discussion of content moderation and the level to which it should be used would stunt progress in any direction. That being said, the most notable finding from this analysis is that we should not strive to choose between having moderation or not having moderation, but instead to recognize that there are certain circumstances where the danger to public safety is so great that we must err on the side of caution. In other words, it is not likely that a blanket solution will ever be sufficient to address the concerns from either stance on moderation; approaching such concerns should be situational and case-by-case. While everyone's satisfaction is not guaranteed, the prioritization of public safety over freedom of speech does ensure that fewer lives are at risk. Nonetheless, preserving freedom of speech in non-threatening situations is imperative, as positive change is initiated by differing thought and dissent. Thus, we must approach content moderation with an expectation of compromise to increase the likelihood of meaningful progress.

Conclusion

The debate around content moderation will be pertinent as long as tools to moderate content change and new components find their place in the content moderation system. That being said, any effort at reformation will be slow as it depends on and competes with the rapid evolution of the moderation algorithms themselves; however, reformation is a necessary struggle and a step in the right direction. Although the pro-moderation and anti-moderation groups yield different goals (public safety and freedom of speech, respectively), their motivations for these goals are both rooted in enhancing the greater good. Meanwhile, their proposed solutions (more

moderation versus less moderation) appear adversarial in nature until a compromise that preserves both goals can be found.

It is important to note that this conclusion has been drawn under three assumptions. First, this paper assumes that the fallacies presented in Noah Carl's paper are wholly true and unbiased. All individuals are subject to implicit biases, and Carl is no different. Consequently, it is understandable that his impartiality in his assertions is questionable; however, all philosophical frameworks are developed from individuals with biases, and Carl's framework is still functional despite his bias. Second, this paper assumes that the evidence I analyzed accurately represents the entire system of content moderation. Given that I was responsible for collection of evidence, any of my personal biases could have impacted the evidence selection process. However, I intentionally sought out evidence that supported both arguments to stifle the impact of my own bias. Finally, this paper begins with the assumption that both free speech and public safety should be valued equally. While this assumption was present in my method, my analysis revealed that the risks posed to societal safety by the system of content moderation had far more negative effects than the risks posed to free speech.

The takeaway from the last assumption came as quite a surprise to me, as I believed the solution to dissent regarding content moderation would be simple – we should either enhance moderation or eradicate it – but I now understand that the solution is far more complex. We must find the Golden Mean, the balance between excess and deficiency in content moderation. This balance is unique, as it does not guarantee that both parties' concerns will be addressed or valued equally, but rather it ensures that one value is prioritized over the other depending on the situation. Content moderation must be situational. My analysis suggests a stronger preference for prioritization of public safety in situations where lives are at risk, for the risk of life will

always outweigh the cost of censorship. This finding is valuable because it addresses the situational nature of content moderation and can help social media companies alter their moderation policies to account for this fact. Content moderation tools must be adapted to favor public safety over freedom of speech in instances where physical safety is compromised.

References

Anti-Defamation League. (2022). Online hate and harassment: The American experience 2021. https://www.adl.org/online-hate-2021

Carl, N. (2018). How stifling debate around race, genes, and IQ can do harm. *Evolutionary Psychological Science*, 4, 399–407. https://doi.org/10.1007/s40806-018-0152-x

Chotiner, I. (2019, July 5). The underworld of online content moderation. The New Yorker. https://www.newyorker.com/news/q-and-a/the-underworld-of-online-content-moderation

Cramer, R. J., Fording, R. C., Gerstenfeld, P., Kehn, A., Marsden, J., Deitle, C., King, A., Smart, S., & Nobles, M. R. (2020). Hate-motivated behavior: Impacts, risk factors, and interventions. *Health Affairs, Health Policy Brief*. https://doi.org/10.1377/hpb20200929.601434

Copia Institute. (2021, August 25). Content moderation case study: YouTube deals with disturbing content disguised as videos for kids (2017). TechDirt. https://www.techdirt.com/2021/08/25/content-moderation-case-study-youtube-deals-with-disturbing-content-disguised-as-videos-kids-2017/

Copia Institute. (2021, November). Twitter briefly restricts account of writer reporting from the West Bank (2021). Trust & Safety Foundation. https://trustandsafetyfoundation.org/blog/twitter-briefly-restricts-account-of-writer-reporting-from-the-west-bank-2021/

Dawson, C. (2020, July 7). Context reduces racial bias in hate speech detection algorithms. ScienceDaily. https://www.sciencedaily.com/releases/2020/07/200707113229.htm

Eidelman, V., Lee, A., & Walter-Johnson, F. (2021, May 17). Time and again, social media giants get content moderation wrong: Silencing speech about Al-Aqsa Mosque is just the latest example. ACLU. https://www.aclu.org/news/free-speech/time-and-again-social-media-giants-get-content-moderation-wrong-silencing-speech-about-al-aqsa-mosque-is-just-the-latest-example

Elliott, D. (2022, August 12). The Charlottesville Rally 5 years later: "It's what you're still trying to forget." National Public Radio. https://www.npr.org/2022/08/12/1116942725/the-charlottesville-rally-5-years-later-its-what-youre-still-trying-to-forget

Facebook. (n.d.). How does Facebook use artificial intelligence to moderate content?. Facebook Help. https://www.facebook.com/help/1584908458516247

Feinburg, A. (2017, December 13). This is the Daily Stormer's playbook. HuffPost. https://www.huffpost.com/entry/daily-stormer-nazi-style-guide_n_5a2ece19e4b0ce3b344492f2

Guthrie, E. (2018, October 29). Gab: Its history and influence in the Tree of Life shooting. The Pitt News. https://pittnews.com/article/137102/news/gab-its-history-and-influence-in-the-tree-of-life-shooting/

Implicit Bias. (2020, January 3). Workplace Strategies for Mental Health. https://www.workplacestrategiesformentalhealth.com/resources/implicit-bias

Katz, A. (2017). Clashes over a show of white nationalism in Charlottesville turn deadly. Time Magazine. https://time.com/charlottesville-white-nationalist-rally-clashes/

Klett, J. (2018, July 20). SCOT. STS infrastructures, Platform for Experimental Collaborative Ethnography. https://stsinfrastructures.org/content/scot

Koch, Dr. A. (2020, October 14). Online white supremacy: Looking for a place to spread hate in the age of multiple communication platforms. GNET. https://gnet-research.org/2020/10/14/online-white-supremacy-looking-for-a-place-to-spread-hate-in-the-age-of-multiple-communication-platforms/

Lavin, T. (2018, January 7). The Neo-Nazis of the Daily Stormer wander the digital wilderness. The New Yorker. https://www.newyorker.com/tech/annals-of-technology/the-neo-nazis-of-the-daily-stormer-wander-the-digital-wilderness

Mac, R. (2021, May 12). Instagram censored posts about one of Islam's holiest Mosques, drawing employee ire. BuzzFeed News. https://www.buzzfeednews.com/article/ryanmac/instagram-facebook-censored-al-aqsa-mosque

McNamee, R. (2020, June 24). Social media platforms claim moderation will reduce harassment, disinformation and conspiracies. It won't. Time Magazine. https://time.com/5855733/social-media-platforms-claim-moderation-will-reduce-harassment-disinformation-and-conspiracies-it-wont/

Nichols, S. (2022, August 16). Cloudflare: We dumped Daily Stormer not because they're Nazis but because they said we love Nazis. The Register. https://www.theregister.com/2017/08/16/cloudflare_ceo_daily_stormer/

Pinch, T. J., & Bijker, W. E. (1984). The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science*, 14(3), 399–441.

Prince, M. (2017, August 16). Why we terminated Daily Stormer. Cloudflare. https://blog.cloudflare.com/why-we-terminated-daily-stormer/

Ricee, S. (2021, May 26). Subconscious vs unconscious: The complete comparison. Diversity for Social Impact. https://diversity.social/unconscious-vs-subconscious/

Torba, A. [@a]. (2021, July 1). Gab is a First Amendment company which means we tolerate "offensive" but legal speech [Gab Post]. Gab. https://gab.com/a/posts/106508069363422579

Townsend, C. (2018, November 29). ElsaGate: The problem with algorithms. United States Cybersecurity Magazine. https://www.uscybersecurity.net/elsagate/