THE MANIPULATION OF FREE SPEECH: AN EXPLORATION OF HOW ONLINE PLATFORMS ENGAGE 'FREE SPEECH' TO SILENCE FREE SPEECH

A Dissertation
Presented to
The Academic Faculty

By

Pooja Casula

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Literature, Media, and Communication
Digital Media

Georgia Institute of Technology

May 2025

THE MANIPULATION OF FREE SPEECH: AN EXPLORATION OF HOW ONLINE PLATFORMS ENGAGE 'FREE SPEECH' TO SILENCE FREE SPEECH

Thesis committee:

Dr. Richmond Wong, Advisor
School of Literature, Media,
and Communication
Georgia Institute of Technology

Dr. Carol Colatrella
School of Literature, Media,
and Communication
Georgia Institute of Technology

Dr. André Brock
School of Literature, Media,
and Communication
Georgia Institute of Technology

Dr. Anne Sullivan
School of the Arts, Media,
Performance, and Design
York University

Dr. Dhanaraj Thakur
Research Director
Center for Democracy & Technology

Date approved: Apr 11, 2025

Free speech is the whole thing, the whole ball game. Free speech is life itself.

Salman Rushdie

ACKNOWLEDGMENTS

The last five years have been an incredible journey, and it is my good fortune that I've been surrounded by the most wonderful people during it. A lot of major thank-yous are in order.

I'd like to first thank my advisor, Dr. Richmond Wong. I have learned so much in the past three years of working with you. Your mentorship, consistent encouragement, and belief in my skills has enabled me to grow so much as a scholar and, more importantly, as a person. So thank you, thank you, thank you, for taking a chance on a third-year PhD student and agreeing to advise her.

To my committee, Dr. Anne Sullivan, Prof. Carol Colatrella, Prof. Andre Brock, and Dr. Dhanaraj Thakur - Thank you so much for your support, advice, and feedback over these last few years! Your guidance has helped me develop as a scholar, and it has been an absolute privilege learning from you all.

To my Therapy Four Group, Shubhangi, Mohsin, and Sylvia. Wow, what a rollercoaster it has been, but I am so glad that you all were there and along for the ride. Shubhangi, thank you for always encouraging me and pushing me to be optimistic and to expect more. The path of my PhD would have looked very different had you not been there, helping me along the way. Mohsin, thank you for your unwavering belief in my abilities and for always making the time for the countless coffee discussions that have doubled as therapy. I owe you many. And Sylvia, thank you for offering sound advice whenever I've needed it, hosting fantastic pizza nights, and for every piece of feedback you've given on my writing. This dissertation is better for it. This PhD would not have happened if it weren't for you three. So thank you, for always being there.

To Allie, Hannah, Inha, Sarah M., Jasmine, Blaire & Rachel - Thank you all for making this PhD a supportive experience. Community is fundamental when you're in grad school, and I am incredibly lucky to have had a community so caring and positive. So thank you

for the coffees, bubble teas, hot yoga sessions, lunches, dinners, and random hallway chats that not only gave me joy but also inspired and encouraged me.

To Sarah B., Harshini, Kavya, Neha, Bharath, Becky, and Miriam - Thank you all for your encouragement, reassurance, and patience these past 5 years. Thank you for always providing the space to talk and for always listening to my rants, research ideas, and inane discussions despite how boring they may be. Also, a special shoutout to Sarah B. and Harshini, my fellow "useless doctors" who are also graduating this year.

To all the participants who enthusiastically joined my studies - Without you, this dissertation would not even exist! Thank you all for your time, patience, interest, and enthusiasm regarding this research.

To my internship mentors from both Amazon and the Center for Democracy & Technology, Sujit, Sheel, Kate, Tanner, DeVan, and Dhanaraj - The internships that I did throughout the duration of my PhD have greatly affected my thinking, methodologies, and framing of research. So, thank you so much for your time, support, and advice during these summer internships!

And finally, to my parents - Everything that I am is because of you. So thank you for being there, for listening, and just in general, for everything.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This dissertation examines the underlying interpretation of 'free speech' in social media platforms and its influence in shaping engagement in online political discourse among young adult Gen Z women. I begin this work by drawing on Legal, Philosophical, and Computer-Mediated Communication (CMC) scholarship to examine dominant theories of free speech in the United States and their influence on social media practices. I then explore how young adult Gen Z women engage in public political discourse online, focusing on their understandings of free speech and how they follow and or differ from the culturally pervasive interpretation of 'free speech'. Following this, I employ my findings and previous theoretical explorations as a lens to analyze online speech moderation policies.

Through this work, I find that social media platforms engage the pervasive US-interpretation of 'free speech' to justify the implementation of moderation policies that excuse toxic behaviors. The continued existence of toxic behaviors online negatively affects how young adult Gen Z women engage in these public discourse spaces, deterring them from publicly engaging and leading them to self-censor.

This raises the question, by engaging the US-interpretation of free speech in their policies and practices, are social media platforms actually suppressing free speech? To address this question, I engage the findings from my previous explorations to investigate the potential of engaging an alternative, relational theory of free speech in platform policies and how this may alter young adult Gen Z women's forms of engagement in online political discourse. I draw on design fiction to re-conceptualize moderation practices rooted in a relational interpretation of free speech and ideate on how this underlying change may affect engagement in public spaces of discourse. Through this work, I hope to encourage researchers and scholars to question and challenge how the culturally pervasive US-interpretation of free speech is being engaged by social media platforms. Specifically, to question whether it is being used to silence rather than to enable free speech.

CHAPTER 1
INTRODUCTION

The concept of free speech has historically been central to the sociopolitical development of the United States. Early U.S. political leaders such as James Madison and Thomas Jefferson indicated that free speech, for land-owning white men, was necessary for the development and existence of a democratic society. Without the ability to freely express one's self and thoughts without ramifications, a democratic society would not function [1]. The First Amendment of the U.S. Constitution states, "Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press..." [1]. The First Amendment protects individuals in the U.S. from the government infringing upon specific individual rights, one of which is the right to free speech. However, the influence of 'free speech' as a concept extends beyond the government and law. Online, 'free speech' is often discussed, debated, and used to justify specific platform practices and policies [2, 3, 4, 5], despite the term originating in the context of the government. This dissertation broadly explores the concept and influence of 'free speech' in the context of social media in the United States.

Although 'free speech' does not have one definition, there is one that has been widely accepted in the United States. John Stuart Mill's 'truth theory' of free speech has shaped the U.S. understanding of what is 'free speech' [6]. In his book On Liberty, Mill declares that all speech should be protected and there should be minimal censorship, arguing that "If all mankind minus one were of one opinion, mankind would be no more justified in silencing that one person than he, if he had the power, would be justified in silencing mankind' [7]. Mill's 'truth theory' states that freedom of speech should be protected because it is vital to the discovery of truth in a society. A key component of Mill's truth theory is the 'marketplace of ideas' concept. The 'marketplace of ideas', a term coined by U.S. Supreme Court

Justice Oliver Wendell Holmes Jr., is based on capitalist laissez-faire principles, explaining that there is a minimal need for censorship as in the 'marketplace of ideas', the truth will emerge only through free and unregulated debate [8, 9, 10].

The 'truth theory' of free speech and the 'marketplace of ideas' concept has repeatedly been quoted and cited by U.S. Supreme Court Justices, in addition to U.S. politicians and business leaders when discussing issues of free speech and expression [11, 6]. In recent years, the term 'marketplace' has often been swapped for 'town square' when discussing social media platforms [9]. This term 'town square' is also a reference to Habermas's concept of the 'public sphere', in which people would gather to discuss and debate [12]. Similar to Mill, Habermas advocated for open discussion and debate in the public sphere as this would be the only way for democratic processes to exist [13]. Elon Musk, the current owner of the social media platform X, has referred to the platform as a 'digital town square' drawing attention to the platform's status as a space for open public discourse [14]. In a 2019 essay, Meta CEO Mark Zuckerberg similarly refers to both Facebook and Instagram as the 'digital equivalent of a town square' [15].

As of 2025, social media platforms such as Facebook and X are considered important spaces of public discourse, given their size and influence [16]. Since their beginnings, these platforms have facilitated sociopolitical movements, influenced elections, and served as important political tools [17, 18, 19]. Due to their size and influence, platform practices and policies shape how users engage in public discourse [16]. This raises the question of whether Mill's theory and the US-interpretation of 'free speech' remains valid when much of the public speech that occurs on social media is viral and anonymous [20, 21, 22, 23]. Instances of mis/disinformation and hate speech have not only impacted one's ability to speak freely in online public spaces of discourse, but they have significantly negatively affected democratic processes around the world [24]. For instance, researchers found that both the 2016 and 2020 U.S. Presidential elections were largely impacted by mis/disinformation surrounding the candidates and voting mechanisms [25, 26]. Political candidates have cited

unmitigated hateful and threatening speech as reasons for dropping their campaigns and or choosing not to run for public office [27]. In a 2018 report, the UN noted that unmitigated hateful and threatening online speech targeting women in public-facing positions broadly affects political participation among women, and may lead to a 'democratic deficit' [20]. Although both mis/disinformation and hate speech are considered 'free' under the First Amendment [28], their existence online has negatively affected both online and offline forms of public political discourse. Since social media platforms hold the power to shape public discourse, and there have been prior instances in which platform practices have negatively impacted public discourse, one has reason to question the platform's interpretation of free speech.

Social media platforms such as Facebook, Instagram, Twitter (now X), and TikTok have, in the past, taken action to address hateful speech and mis/disinformation, particularly concerning political and election processes. Many of these measures include removing speech from platforms, banning or suspending users from platforms, or in some cases, limiting their reach on platforms (shadowbanning) [16]. There has been significant pushback regarding whether these measures to limit mis/disinformation and hate speech have served to stifle 'free speech' in the United States. Many free speech absolutists argue that these measures are a form of censorship, as hate speech and mis/disinformation are still considered 'free' under the First Amendment, and restricting these forms of speech would therefore be in opposition to the original interpretation of 'free speech' [29, 30, 31].

It is important to note that the First Amendment protects citizens from the U.S. government infringing upon their free speech rights. Companies, and therefore social media platforms, do not need to adhere to the First Amendment, and by proxy, its interpretation of 'free speech'. However, as stated earlier, this US-interpretation of 'free speech' is pervasive and has influence beyond the Constitution. Despite not needing to adhere to these interpretations of 'free speech', companies continue to do so. At the time of writing, both Meta and X have moved to limit and or completely remove several misinformation and hateful

speech moderation measures, citing a desire to increase 'free speech' on the platform [32, 33, 34].

The continued influence of free speech absolutism online raises concerns regarding public perceptions of how political discourse should occur. As previously noted, the practices of social media platforms 'shape the shape of public discourse' through their decisions of what content and discourse is allowed on the platform [16]. For instance, a decision to not moderate misogynistic disinformation targeting a political candidate plays a role in shaping public perceptions of whether political disinformation is legitimate political discourse and therefore 'free speech'. This is particularly true for younger social media users, including young adults and children, who have been exposed to social media platforms and their culture for the majority of their lives. These users' exposure to platform practices has shaped their general perceptions of 'free speech' and what constitutes political discourse.

In this dissertation, I explore how underlying conceptions of 'free speech' in platform moderation practices may broadly shape user perspectives of social media public discourse. I aim to understand what content social media users perceive should be 'free', how their perspectives of 'free speech' shape their patterns of engagement on the platform, and how introducing an alternative interpretation of 'free speech' underlying platform practices may affect their perspective of engaging in public political discourse online. Specifically, I examine these questions through the experiential lens of young adult Gen Z women.

1.1 Motivation

While existing research in HCI and social computing has closely examined online engagement, political participation, and content moderation practices, much of this work has broadly not engaged with free speech theory. Much of this work has focused on mechanisms to improve upon moderation practices [35], such as through automated methods [36, 37] or crowdsourced methods [38]. Researchers have also developed frameworks for moderation on platforms [39] and have done work to understand how users employ plat-

forms in election processes and or activism [17, 18, 40]. All of this work is essential to better understand the larger ecosystem in which social media, political processes, humans, and businesses co-exist and are connected. However, in parallel to this work, there has been much scholarship in law and political science grappling with some of the same questions from an alternative lens.

In U.S. law and political science, questions surrounding online moderation and online political participation are often discussed within the larger context of free speech and the First Amendment. Legal scholars have written about the legality and consequences of moderating speech in online contexts concerning the First Amendment [41, 4, 42]. Philosophers, social scientists, and sociolinguists have written about how companies should be discussing and handling the moderation of speech, in addition to how the moderation of speech impacts public perceptions [43, 44, 45, 46], particularly when 'free speech' is such a central component of US culture and ideology. These explorations of the politics of platforms and content moderation are equally as necessary to better understand the influence that platform practices and policies have on shaping public perceptions of free speech in politics and broader sociopolitical discourse.

Young adults, specifically those among the Gen Z generation (born 1997 - 2012) [47, 48], have been exposed to social media culture and environments for a significant part of their lives, shaping their perspective of how online environments may influence political processes [49, 50]. The eldest members of Gen Z (born 1997-2000) were in elementary school during the 2008 U.S. Presidential election and observed the normalization of political candidates turning to social media platforms such as Twitter and Facebook to disseminate political messaging [51]. They were in school during the 2011 Arab Spring, observing the ability of platforms like Twitter to mobilize and aid activist efforts [52, 53]. Members of Gen Z were able to vote during the 2016 and 2020 U.S. Presidential elections, joining the broader political discourse at a time when online terms such as 'mis/disinformation', and 'echo chambers' became more widespread and common offline [25, 54, 55]. One may

argue that members of Gen Z have joined the broader U.S. political discourse at a critical juncture, where questions have been raised regarding the validity of the entire political system. Their perspective of the sociopolitical environment has largely been shaped by their exposure to social media discourse.

I narrow the scope of this work to focus on young adult Gen Z women due to their unique experience in spaces of political discourse. Prior work in political science, CMC, HCI, and social computing has examined how users who identify as women are often excluded and or marginalized in online spaces of discourse, shaping their forms of engagement [56, 57, 58]. Studies have found a prominent gender gap in public forms of engagement, such as posting and commenting in online political and news spaces, as many women feel discomfort and unsafe openly participating in these spaces [59, 60]. The fear of online abuse, harassment, doxxing, offline violence, and stalking all serve to deter women from publicly participating in spaces of discourse online [61, 62]. In her book, Misogyny Online, Emma Jane notes how many women who want to hold social media accounts will deliberately hide aspects of their personal life and gender identity when posting on the Internet, concealing or changing their name, and or avoiding any discussion of activities or issues relating to women [63]. This behavior indicates that these users hide or change their identity in order to participate in these public spaces. These forms of engagement raise the question of whether women feel that they are enabled to freely and openly express themselves in public online spaces without ramifications. Or, has their ability to engage in free speech been compromised?

The right to free speech has often been used to justify the protection of violent, threatening, and offensive speech. As the American Civil Liberties Union (ACLU) states, in order to protect one's right to free speech, we must protect 'the thought that we hate' [64, 65]. While many have argued against this 'slippery slope' logic [66], it remains the most commonly used reasoning to protect violent, threatening, and offensive speech. However, when such speech is permitted in spaces of public discourse, it often leads to a significant silenc-

ing effect in which others feel the need to self-censor or limit their own speech out of fear of facing ramifications for expressing their opinions [67, 68, 61]. At a time in the United States when there are concerns regarding one's ability to freely express their opinions on justice and equity without experiencing professional and personal ramifications [69], I raise the question, should platforms be employing the US-interpretation of free speech? If our online spaces of discourse, which do not need to be beholden to the US-interpretation of free speech, cannot provide us the space to freely express our opinions, what might we do? How might we challenge the interpretation of what is 'free speech' in these spaces? Positioning myself as a scholar in the social computing/CSCW space, I draw on legal, political science, and CMC theories and scholarship to inform my exploration of these questions.

1.2 Research Questions

There are three primary research questions motivating this research.

> **(RQ1) How do Gen Z young adult women perceive and engage in social media spaces of political discourse?**
>
> **(RQ2) What interpretations of 'free speech' underlie social media speech moderation policies, and how do they shape moderation practices?**
>
> **(RQ3) How might engaging a relational theory of 'free speech' in social media speech moderation policies affect engagement in online political discourse spaces among young adult Gen Z women?**

1.3 Dissertation Overview

1.3.1 Chapter 2

In Chapter 2, I present existing literature on free speech in the United States, content moderation on social media platforms, and political participation on social media. I begin by

presenting two prominent theories of free speech in the United States, the truth theory and the self-realization/autonomy theory, and their influence on the US-interpretation of free speech. I then explore prior literature on free speech on the Internet, including perceptions of and practices of content moderation. Following this, I present existing literature in HCI and political science on political participation among youth and women online. In exploring this prior work, I develop an understanding of what is the 'US-interpretation of free speech' and its influence on Internet culture and practices.

1.3.2 Chapter 3

In Chapter 3, I investigate RQ1: How do Gen Z young adult women perceive and engage in social media spaces of political discourse? To do this, I conduct an interview and diary study with young adult Gen Z women (ages 18 - 27) to understand their perceptions towards and forms of engagement in spaces of online political discourse. The findings of both studies indicate that young adult Gen Z women are broadly critical of social media being used as a space to discuss politics, are skeptical of the effectiveness of using social media for political purposes, and generally disengage from online spaces of political discourse. Drawing on the pervasive US-interpretation of free speech as a lens, I examine their perceptions and forms of engagement and conclude that the young adult Gen Z women participants in these studies hold an alternative conception of free speech than the one that is prevalent on social media. Specifically, their conception of free speech emphasizes understanding the situated context of the speech, and the relationship between the parties involved in the speech.

1.3.3 Chapter 4

Chapter 4 broadly investigates RQ2: What interpretations of 'free speech' underlie social media speech moderation policies, and how do they shape moderation practices? In this chapter, I present two close readings of speech moderation policies to examine the un-

derlying notions of free speech. I present a close reading of X's Violent Content Policy (as of Oct 2024) [70], and Meta's Violence and Incitement policy (as of Oct 2024) [71]. Through my close readings of both policies, I draw attention to the underlying Libertarian notions of free speech and discuss the potential implications of engaging these notions of free speech in platform policies on forms of engagement among young adult Gen Z women in online political discourse. I argue that these platform policies draw on free speech absolutism, a large influence on the US-interpretation of free speech, to center speakers and their intentions over the reaction of targeted parties involved in contentious online posts. This speaker-centrism may contribute to potential chilling effects on the intended audience. While these policies, and the US-interpretation of free speech, are concerned regarding potential chilling effects on speakers/posters, they disregard the potential for chilling effects on those who have been targeted by speakers/posters.

1.3.4 Chapter 5

In Chapter 5, I build on the findings from previous chapters to explore how engaging a relational theory of free speech in online speech moderation policies may affect how young adult Gen Z women engage in spaces of online political discourse. Specifically, I ask, (RQ3) How might engaging a relational theory of 'free speech' in social media speech moderation policies affect engagement in online political discourse spaces among young adult Gen Z women? Through an ideative workshop study with 9 young adult Gen Z women, I found that participants advocated for mechanisms for negotiation, reflection, and more human intervention in a moderation system rooted in a relational theory of free speech. However, I also found that participants felt that changes made to moderation systems would largely not affect their practices of engagement in online spaces of political discourse. Their current perceptions of 'free speech' and moderation limited their ability to question and challenge what moderation could be under an alternative 'free speech'. Many believed that the underlying culture and practices of platforms needed to change substan-

tially for them to alter their forms of engagement in political discourse spaces. Understanding that the workshop study did not effectively explain how significantly platform practices and policies may change under a relational theory of free speech, I created a design fiction to highlight the potential of this change.

1.3.5 Chapter 6

In Chapter 6, I use design fiction as a thought exercise to explore how content moderation systems may be designed with relational theories of free speech. The design fiction aims to draw attention to the potential of implementing an online speech moderation system rooted in an alternative interpretation of free speech that centers on understanding the relationship between a speaker and their audience and the situated contexts in which the speech was made. I present the fictional moderation service through a guide on how to use the service for both users and moderators, in addition to a set of example posts. Through these fictional posts, I discuss both the potential and the limitations of engaging a relational interpretation of free speech in the design of online content moderation policies. Furthermore, I compare and contrast components of the US-interpretation of free speech and the relational theory of free speech in the context of online moderation. Specifically, I focus on the differences in how these theories interpret intentionality, credibility, the relationship between speakers and audiences, and in how these theories interpret chilling effects. Through this design fiction exercise, I highlight the potential of engaging an alternative theory of free speech in platform moderation practices to impact broader experiences in online public spaces of political discourse.

1.3.6 Chapter 7

Chapter 7 concludes this dissertation by discussing the current state of moderation and social media platform practices in the context of political discourse. As the second Trump administration takes office, social media platforms have made several significant changes

to their moderation policies, citing a desire to increase 'free speech' [32]. These changes have re-ignited debate surrounding free speech, mis/disinformation, hate speech, and political discourse. These changes also remind social media users that moderation policies are dynamic and what may be considered 'free speech' one day, may not be the next day. As the concept of 'free speech' is manipulated by platforms to justify boosting engagement through hateful rhetoric and generating a silencing effect, I discuss the importance of researchers seeking and exploring alternative approaches to moderation rooted in alternative interpretations of free speech.

CHAPTER 2
BACKGROUND & RELATED WORK

This chapter presents existing theoretical and empirical research in the areas that motivate this work. I begin by exploring theories of free speech that have been influential in shaping the US notion of 'freedom of speech' and First Amendment doctrine. I investigate how this US-interpretation of free speech influenced the design of the early Internet and social media platforms. Following this, I explore the state of content moderation on social media, in addition to political participation and discourse online, specifically among young adults and women.

Content Warning: This chapter contains references to violent, hateful, and discriminatory language. I ask that readers please be advised when reading this chapter.

2.1 Free Speech in the United States

"Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances."
- First Amendment, United States Constitution [1]

'Freedom of Speech' has long been considered a fundamental component of the United States democratic system. An individual's right to free speech and free expression is protected by the First Amendment of the United States Constitution, however, there is a wide berth to interpreting what exactly 'free speech' means.

At present, there are eight forms of "speech" not legally protected by the First Amendment and therefore not considered 'free speech'. It is important to note that "speech" under the First Amendment is understood as verbal language, written speech, non-verbal forms of language such as sign language, physical actions made in protest such as not saluting

the flag, in addition to other forms of expression such as clothing [72, 73]. The eight unprotected forms of speech include speech that incites imminent lawless action, true threats, fighting words, obscenity, defamation, fraud, child pornography, and speech integral to criminal conduct [28].

I note that hate speech, such as white supremacist or neo-Nazi speech or symbols, is protected under the First Amendment. The U.S. is widely tolerant of hateful and abusive speech, particularly when compared to other countries like Germany [74]. This laissez-faire attitude towards hateful speech can be traced back to formative court cases regarding the First Amendment. In his dissenting opinion in the 1929 case *United States v. Schwimmer*, Supreme Court Justice Oliver Wendell Holmes Jr. states "... but if there is any principle of the Constitution that more imperatively calls for attachment than any other it is the principle of free thought – not free thought for those who agree with us but freedom for the thought that we hate" [75].

This phrase, "freedom for the thought that we hate" has been adopted and adjusted by various institutions in the United States, including the American Civil Liberties Union, as the primary justification for why certain forms of hateful, abusive, and or violent speech should continually be tolerated [65, 66]. Though initially attributed to Justice Holmes, the phrase traces its roots back to English philosopher John Stuart Mill's truth model of free speech. Mill's notion of free speech argues that speech should be protected as it is central to the discovery of truth, and as such, all speech other than threatening speech should be protected [7]. This theory of free speech, the truth model, is further explored below. At present, Mill's model of free speech remains pervasive in courts of law across the United States, in addition to the Internet [76, 77, 9].

Mill's model of 'free speech' has largely influenced the design of online platforms dedicated to public discourse. Early internet researchers have examined the underlying Libertarian ethos of internet culture, particularly the 'technolibertarianism' view, which advocated for minimal regulations on speech and individualism among other things [78].

Present-day social media platforms draw on Mill's interpretation of speech in their platform policies and use this notion of 'free speech' to justify or validate the implementation of these policies [14, 79].

However, many scholars have since questioned the assumptions and dimensions of the Western individualist interpretation of 'free speech', specifically concerning whose speech this understanding of 'free speech' is protecting [80]. As U.S. legal scholar Mary Anne Franks notes, "Far from being a universal, objective doctrine, First Amendment law has a gender, and a race, and a class" [66]. This culturally pervasive understanding of free speech centers only on the ability of the speaker to exercise 'free speech'. Alternative notions of free speech, such as feminist free speech theories and relational free speech theories, center the roles of situatedness and context of both the speaker and the listener during communication [66, 76].

In the following section, I examine two theories of free speech that have shaped the interpretation of free speech that is widely pervasive in the United States. The two theories, the truth theory, and the autonomy theory advocate for minimal regulations on speech but differ in their justification for why speech should be protected. The truth theory is predicated on the belief that speech should be protected because it is imperative to discover the truth. The autonomy theory is rooted in the belief that speech should be protected because speech is a mechanism by which individuals exercise and maintain their autonomy. I elaborate further on these theories of free speech in the following section.

2.2 Dominant Theories of Free Speech in the United States

2.2.1 The Truth Model

The truth model of free speech is perhaps the most influential theory in the development of First Amendment doctrine [81]. This model of free speech, also known as the 'truth theory' was primarily developed by English philosopher John Stuart Mill, building off of the work of another English philosopher, John Milton. In his book On Liberty, Mill

defends the ability to speak freely going as far as to say, “If all mankind minus one were of one opinion, and only one person were of the contrary opinion, mankind would be no more justified in silencing that one person than he, if he had the power, would be justified in silencing mankind” [7]. At its core, Mill’s theory states that it is important to protect free speech because free speech is essential to the discovery of truth. Mill’s definition of truth is rooted in Cartesianism [76], which views truth as an idea that is discoverable and not necessarily something that can be shaped by one’s experience. In Mill’s theory, without free speech, society will never know what the truth is, and therefore, for the benefit of society, we must protect free speech.

To further explain this, Mill proposes the ’marketplace of ideas’ concept. The term ‘marketplace of ideas’ was first used by Supreme Court Justice Oliver Wendell Holmes Jr. [8, 10], however, the concept originates from Mill. Mill argues for the ‘free trade of ideas’, stating that if society were analogous to a marketplace and there is a competition of ideas, then through debate and discussion, the truth would emerge. The truth theory draws largely from capitalistic economic theory [82]. If there is competition between vendors, they will continuously refine their products to ensure they are competing on the same level as their competitors, and thus the best product will emerge. In the case of free speech, Mill states that competition of opinions and ideas will lead to the “truth” emerging. He further argues that censorship, specifically by the government, is not needed in such a situation, as the debate, and thus competition, will ensure that ‘falsehoods will not prevail’ [7].

In fact, the only instance which Mill argues for government interference to censor speech is if the speech in question causes harm to another individual. This is known as the ’Harm Principle’. Mill explains that ’harm’ in this context refers to an action in which one party injures or violates the rights of another party. However, although Mill argues that harmful speech should be mitigated by the government, he does not provide a threshold for which an action is defined as ‘harmful’ [83]. Regardless, the Harm Principle has been used to create tests and standards to determine what speech is permissible and not permissible

under the First Amendment in the United States [84].

Influence on US-Interpretation of Free Speech

John Stuart Mill's truth theory remains influential in the development of First Amendment doctrine. Mill's 'marketplace of ideas' concept has been repeatedly cited in various Supreme Court opinions, and the Harm Principle has been pivotal in the development of what constitutes protected speech [8].

An illustrative example of the truth theory being a major part of First Amendment doctrine is the Supreme Court's decision in the 2011 case *Snyder v. Phelps*. Fred Phelps and other members of the Westboro Baptist Church protested at the funeral of Marine Lance Corporal Matthew Snyder, carrying signs with various hateful slogans including but not limited to, "Thank God for Dead Soldiers", "God Hates Fags", "You're going to hell", and "God Hates You". Members of the Westboro Baptist Church subscribe to the belief that God is punishing the United States by killing soldiers due to the country's tolerance of the LGBTQIA+ community, and to perpetuate this belief they often stage protests at military funerals. Members of the Church protested relatively far away from the specific funeral location and remained non-violent the entire time. The Snyder family filed a lawsuit against Phelps and the Church, one of the charges being 'intentional infliction of emotional distress'.

In an 8-1 decision, the Supreme Court ruled in favor of the Westboro Baptist Church, stating that their speech, via their signage, was within their First Amendment rights. The Court argued that the First Amendment protects speech that is a matter of public concern, and the content of the protestors' signs was general and referring to a matter of public concern, especially because they had protested at other military funerals as well. In other words, the protestors were not directly targeting Marine Lance Corporal Matthew Snyder, they were sharing their general beliefs on the matter. Therefore, their conduct cannot be understood as intentionally inflicting emotional distress on the Snyder family.

On behalf of the Court, Chief Justice Roberts states in his opinion, “Speech is powerful. It can stir people to action, move them to tears of both joy and sorrow and —as it did here—inflict great pain. On the facts before us, we cannot react to that pain by punishing the speaker. As a Nation we have chosen a different course —to protect even hurtful speech on public issues to ensure that we do not stifle public debate” [85].

Chief Justice Robert’s words indicate that the interpretation of free speech in the United States includes the protection of speech that is hateful and hurtful, regardless of its effects on the listener or audience. For instance, even if the speech were to cause a listener to self-censor out of fear, it would remain protected under this interpretation of free speech. Such an interpretation appears to favor the speaker over the listener or audience of the speech.

2.2.2 Autonomy, Self-Realization, & Liberty Theory

The autonomy theory of free speech is predicated on the assumption that speech should be protected because it is fundamental to maintaining the autonomy and ‘self-fulfillment’ of an individual. Legal scholar Susan Williams explains that the autonomy theory is a ‘family of theories relating to self-realization, dignity, choice, and the social contract’ [76]. Due to this, there are various similar sub-versions of this larger theory that highlight different dimensions of autonomy. Throughout this dissertation, I will refer to this family of theories by the terms ’autonomy theory’, ’self-realization theory’, and or ’liberty theory’.

One version of autonomy theory states that speech should be protected because it is ‘instrumental’ to one’s self-realization and to exercise their autonomy. Per this version of autonomy theory, speech enables one to question and challenge, increasing their autonomy, and therefore should be protected. Similarly, a few theories of speech that emphasize autonomy, such as legal scholar C. Edwin Baker’s liberty theory, argue that speech should be protected because it enables one’s self-realization and liberty [81, 86]. In his book, Human Liberty and Freedom of Speech, Baker begins by arguing, “The liberty model holds that the free speech clause protects not a marketplace, but rather an arena of

individual liberty from certain types of government restrictions" [81].

Baker argues that Mill's 'marketplace of ideas' concept makes three unreasonable assumptions: 1) there is such a thing as 'objective truth' and it can emerge from competitive debate, 2) people are 'rational' and are able to discern what is true and what isn't, and 3) people want to know the truth [81]. Baker argues that the concept of 'objective truth' is unproven and that people's perceptions of what is 'true' vary depending on the lives that they have lived. Thus, what an individual may perceive as 'true', another may not, and Mill's theory doesn't account for this. Baker further argues that people have differing agendas, and they may not necessarily want to engage in debate to find the 'truth' if it doesn't mean anything to them or further their agenda [81].

Baker rejects the idea that freedom of speech should be protected because it is necessary for the benefit of society as a whole. Instead, he believes that free speech should be protected because it holds value to the individual as a mechanism to maintain one's liberty and autonomy. Baker illustrates this through an example of a war protester, arguing that their speech may not necessarily have an impact on the war effort, but by using speech to protest, the protester is defining themselves and their stance. In this case, speech is used as a mechanism of self-definition, and by proxy, self-realization. These interpretations of free speech and speech are individualistic, emphasizing the value and role of free speech to the individual.

Unlike Mill, Baker does not advocate for absolute free speech. Baker argues that threatening speech, speech forced on an individual (i.e. something an individual was forced, required, or coerced to speak), espionage, and treasonous speech should all not be protected under free speech. However, Baker notes that obscene or hateful speech should be protected [87].

Influence on US-Interpretation of Free Speech

The self-realization and self-fulfillment interpretations of free speech are foundational to the Libertarian movement and interpretation of free speech. They emphasize individual liberties and rights. These theories are also widely cited in U.S. legal scholarship and law. Many Supreme Court opinions engage with these perspectives and interpretations of speech and free speech, such as in the cases of *West Virginia State Board of Education v. Barnette* (1943) [88], *Stanley v. Georgia* (1969) [89], and *303 Creative LLC. v. Elenis* (2023) [90].

In the 2023 case *303 Creative LLC. v. Elenis*, the Court considered the question of whether the Colorado Anti-Discrimination Act violated the First Amendment rights of a wedding website designer because it did not allow her to publish a statement declining to design wedding websites for same-sex weddings when she opposes them on religious grounds. In this case, Lorie Smith, on behalf of her company 303 Creative LLC. challenged Colorado's law, arguing that it violated her individual right to free expression to publish her statement regarding her opposition to same-sex marriage. The law prohibits a business from discriminating against clients based on 'protected characteristics' such as race, sexual identity, etc. In a 6-3 decision, the Court ruled that the Colorado law did violate Lorie Smith's rights to free expression, as the law compelled her to create speech she did not believe in. Justice Neil Gorsuch delivered the Opinion of the Court stating, "As surely as Ms. Smith seeks to engage in protected First Amendment speech, Colorado seeks to compel speech Ms. Smith does not wish to provide...Here, Colorado seeks to put Ms. Smith to a similar choice: If she wishes to speak, she must either speak as the State demands or face sanctions for expressing her own beliefs…" [90].

Gorsuch's Opinion invokes the autonomy interpretation of free expression, arguing that speech should be protected due to its value to the individual and exercising their right to maintain their autonomy and not yield to governmental intervention. Gorsuch later states, "...the First Amendment protects an individual's right to speak his mind regardless of whether the government considers his speech sensible and well-intentioned or deeply

"misguided," and likely to cause "anguish" or "incalculable grief", echoing the interpretations' emphasis on protecting speech that is hateful regardless of its impact on its intended audience.

It is important to note that while these theories of speech vary in their interpretation of why speech should be protected, a combination of both theories is often used by the Courts, as both theories contain similar elements. For instance, both theories call for minimal censorship. The 'free speech absolutism' ideology, which pushes for minimal censorship, is deeply rooted in both Mill's truth theory and theories of autonomy and self-realization. Legal scholarship and First Amendment doctrine do not strictly abide by one theory or another but rather draw on specific dimensions of each theory when implementing them. Furthermore, each Supreme Court Justice may personally abide by differing interpretations of why free speech should be protected, impacting which interpretation is invoked in a Court Opinion. However, these two theories of free speech, the truth model and the self-realization/autonomy model are influential in the wider understandings and interpretations of speech in the United States, specifically those that are online. Throughout this dissertation, I refer to the combination of both theories of free speech as the 'US-interpretation of free speech'. I also use the phrases 'U.S. notions of free speech' and 'free speech absolutism' to describe the culturally pervasive interpretation of free speech in the United States.

2.3 Free Speech and the Internet

The notions of free speech and expression in the United States extend to the Internet. Many researchers have explored the Libertarian ethos underlying early Internet culture, particularly concerning speech [91, 44, 92]. Terms such as "technolibertarianism" are used to describe the utopian, idealistic notion that the Internet should remain unregulated and that all speech should be allowed and encouraged in virtual space, echoing Mill's truth theory and the autonomy theory of free speech [78].

Habermas's concept of the public sphere has also largely shaped perceptions of the Internet. Habermas argues that the 'public sphere' is created for debate and discussion, rather than Mill's 'marketplace' which, he argues, was created to buy and sell [93]. However, Habermas supported Mill's ideology that society requires open, rational discussion and debate to form an opinion and carry out democratic processes [13]. The 'public sphere' is characterized by its separation from governmental, business, and legal entities. In other words, for an institution to be a 'public sphere', it must be separate from the government and the 'market' [93]. Habermas's public sphere is meant to be a space in which anyone can critique their government or corporations without experiencing repercussions. Social media platforms and the Internet at large have often been compared to the 'public sphere', however, since social media platforms are businesses, they cannot be defined as a 'public sphere', despite claiming to be a space dedicated to open debate and discussion [12]. Habermas's 'public sphere' has been considered an ideal that social media platforms and other online spaces have attempted to emulate [13]. However, Feminist scholar Nancy Fraser argues that even Habermas's public sphere is exclusionary, as the spaces Habermas argued were 'public' and open to everyone for discussion were exclusive to white, bourgeois men. Fraser's critique of Habermas highlights the exclusionary practices that underlie the concept of the 'public sphere' [94]. Thus, although Habermas advocates for 'freedom of speech' when conceptualizing the public sphere, the exclusion of various people from the 'public sphere' raises the question, who actually was enabled the 'freedom of speech' in these public spaces? Furthermore, if the ideal that many platforms and discourse spaces are striving to achieve is rooted in exclusionary practices, should it be considered the ideal?

Despite this, research has shown that many Internet users have felt that they can openly speak with others online in a way that they may not be able to offline, in some instances developing a virtual support system and community in the process [95]. However, such spaces do contain certain rules and or social norms that may unofficially regulate behavior. Prior work has shown that spaces that are explicitly unregulated are often exclusionary,

discouraging people who may openly disagree with certain discourses from participating [2, 56, 96]. For instance, in an article examining two instances of sexual harassment online, Linguist Susan Herring specifically notes that the "libertarian principles of freedom of expression" in an online discussion led to any oppositional speech to be thought of as "censorship" and those who made this oppositional speech were kicked out of the space [46].

2.3.1 Section 230

In the United States, in accordance with Section 230 of the Communications Decency Act granted to them by the Federal government, social media platforms are seen as 'distributors' of content, not 'publishers', thereby legally absolving them from facing any consequences for content generated by users. A publisher of content would decide what content to display versus what content not to display. A distributor displays content as it is generated by users and users are responsible for what they post [16, 97, 98]. For instance, X and Facebook are not held responsible for misinformation spread on their platforms because they 'distributed' the content that was generated by their users on the platform, and did not deliberately choose what content to show their users.

Despite early Internet advocates pushing for minimal regulation online, and the existence of Section 230, at present, most social media platforms have created policies and systems to mitigate graphic, violent, hateful, and false content. These regulatory practices are referred to as 'content moderation'. One may argue that Section 230 enables platforms to not moderate content since they are not legally responsible for "third-party" content generated by users on the platform [99]. However, things become more uncertain when content moderation does come into play because although Facebook, X, and Reddit do not necessarily have to moderate user-generated content, they choose to do so. Through these practices, they are determining what content is on the platform, and what is not, therefore toeing the line of being considered "publishers". This change in status has been noted by

the U.S. Congress, and there has been a push for Congress to address Section 230. However, at the time of writing this dissertation, the law continues to exist [100].

2.3.2 Content Moderation

Broadly, content moderation refers to the process of mitigating user-generated content on a platform. Oftentimes, the content being mitigated is harmful, sensitive, threatening, and or violent and is being moderated to avoid further spread across the platform. Methods of content moderation include removing a post, limiting the visibility of a post, banning users, suspending users for a period, and or labeling posts that have specific forms of sensitive or potentially problematic content [16]. In most cases, the content that is reviewed by the system has been flagged or reported by another user [101]. How a platform addresses a specific post is dependent on its content moderation policy.

Platforms vary widely in their approach to content moderation. Some platforms outline specific policies that are applied to the entire platform. For instance, Twitter/X and Facebook have specific policies on how they might handle threats of violence, hateful conduct, election disinformation, child sexual exploitation, and other harmful behaviors that are applied to all users [102, 103]. Other platforms, such as Reddit and Mastadon, in which small sub-communities exist, have specific rules for those sub-communities. While both platforms do have platform-wide rules, for instance, to mitigate child sexual exploitation, much of the moderation takes a decentralized approach, with each sub-community having its own rules for what behavior is acceptable [104, 105, 106].

These policies are enforced in different ways as well. Platforms such as Discord, Facebook, and X enforce their policies via a combination of both human content moderators and automated systems [107, 108, 105]. Platforms such as Reddit and Mastadon rely heavily on human volunteer moderators, who not only enforce moderation rules but also are a part of the sub-community they moderate and who often aid in the creation of the moderation rules [104, 106].

Motivations for Moderating Content

Despite not having much legal incentive for content moderation in the U.S., social media companies have a vested interest in maintaining civility on their platforms. In his treatise, Exit, Voice, and Loyalty, economist Albert O. Hirschman argues that consumers of a product will either 'exit' or 'voice' if they are not provided a service up to a specific standard [109, 110]. In the case of a social media platform, users will either leave the platform or be vocal about their concerns regarding the platform. An example that illustrates this principle is what occurred on Twitter after Elon Musk became the CEO of the company [33]. If users leave the platform, the company will lose money and, therefore will be incentivized to update their service. If users are vocal about their concerns, other users may also become vocal, and this would draw negative attention to the company, potentially prompting it to update their service. In 2015, Twitter's former CEO Dick Costolo acknowledged that the content moderation system at Twitter needed to be better to retain its user base, explicitly stating, "We suck at dealing with abuse and trolls on the platform and we've sucked at it for years. It's no secret and the rest of the world talks about it every day. We lose core user after core user by not addressing simple trolling issues that they face every day" [111]. It is unclear the direction that present-day Twitter/X will take in this situation. As of 2025, the platform has made significant changes to many of its content moderation policies, in addition to its business model as well [112, 113]. Along with these changes, however, is evidence of many users leaving, and other users vocalizing their concerns about the platform [34].

A financial incentive to moderate content is the inclusion of advertising on social media platforms. Prior research has shown that the rise in advertising on social media platforms has incentivized them to remove unsafe and or toxic content, as advertising has become a lucrative stream of revenue [114]. For an advertiser, if a platform contains large amounts of online toxicity and abusive content, they are less likely to promote the product on the platform due to "brand safety issues"[115]. However, research has also shown that social

media platforms are incentivized to remove toxic content only up to a specific threshold, after which it makes no difference in user engagement. Economically, a platform may only choose to remove a specific piece of content if by removing it, user engagement increases and therefore ad revenue also increases [45]. Regardless of the incentive, most social media platforms enact some form of content moderation.

Widespread Effects of Content Moderation

Social media platforms are not the government nor the Supreme Court and, therefore, hold no legal authority over what speech is protected under the First Amendment and what is not. However, their sheer size and influence in some ways provide them the authority to do just that in practice. As Internet scholar Tarleton Gillespie puts it, "Platforms may not shape public discourse by themselves, but they do shape the shape of public discourse" [16]. Platforms shape cultural understandings and societal conceptions of 'free speech'. Many have called social media platforms the 'digital town square' or the online version of the 'marketplace of ideas' where most political discourse occurs [9, 14]. So, when social media platforms, via their content moderation decisions, determine what speech will be seen and heard by the larger audience, they are, in essence, shaping public opinion and understanding [116].

For instance, not mitigating online violent behavior targeting women may be normalizing it. Not mitigating online violent behavior implies that violence targeting women online is 'typical' and sometimes, threatening behavior is just part of the discourse. This is not to say that social media platforms are deliberately ignoring the moderation of online violence. As evident from the previous sections, content moderation is a complex and difficult process. How does a platform mitigate toxicity in a way that enables the free expression of many without suppressing the voices of a few? But given content moderation's role in influencing the shape of public discourse, one has the right to question whether the current ways of approaching content moderation are, in fact, effective.

While much of the current research in the space of content moderation is trying to develop systems that remove hurtful and or violent content more quickly and efficiently, little research examines the underlying philosophy of moderation systems. Specifically, the underlying free expression philosophies within content moderation systems and how they influence the moderation of online behavior. In Chapter 4, I closely examine the underlying free speech interpretations of the speech moderation policies of X and Meta. I explore the influence of the US-interpretation of free speech in the implementation of these policies and how they may shape public discourse online.

Social media spaces are often considered centers of public forms of political discourse, as they provide avenues for public officials to communicate with constituents and for individuals to express their political opinions and ideologies with minimal intervention [117, 118, 119]. Below, I outline prior research examining the role of online spaces in enabling political discourse and participation.

2.4 Online Platforms and Political Participation

Online platforms have long been used as centers of political discourse. Early internet research cited their potential effectiveness in increasing political participation through online discussion, indicative of the fact that online spaces have always been thought of as spaces to discuss and debate [120, 117, 121]. Users turn to online platforms to learn the news, to seek and share information relating to political events, and to engage in discussions regarding current sociopolitical issues [122, 123]. Much of these forms of engagement online occur publicly. Despite prior work examining and questioning the effectiveness of online forms of political engagement to consider it as part of political participation [124, 125, 126], past work in HCI has studied how system and platform design decisions may help users engage publicly in politics online (e.g., [127, 128, 129]).

The literature exploring online users' perceptions of and engagement with political discourse spaces reflects a general apprehension of publicly participating in them, which in

turn significantly affects political participation online and offline. Prior literature on adult and young adult engagement with online political discourse spaces illustrates that an aversion to conflict and argumentation may lead to an avoidance of public political engagement online [130, 18, 131]. Prior literature on women and political discussion spaces online demonstrates that abuse, harassment, and a fear of physical consequences also lead to an avoidance of public political engagement online. Much of this literature highlights the lack of public political engagement online as an indication of a lack of broader political participation [132, 67, 27].

This, in turn, may lead to users seeking out online communities and discourse spaces that share their political values and ideology [133, 134]. Users may utilize their online profiles to signal others of their political leanings [135, 136, 137]. However, research has also shown that the preference for discourse with users holding similar political views has led to the creation of echo chambers, which has a significant effect on furthering political polarization [138, 139, 140].

Much of this prior research does not specifically focus on Gen Z young adult social media users, nor do they examine the role that gender identity potentially plays in young adult online engagement in political spaces. Members of Gen Z have long been exposed to social media and Internet culture, specifically in how it affects politics and political conversations [48]. Furthermore, much of the past work occurred prior to the onset of the COVID-19 pandemic, and as a result, does not reflect the rise in misinformation and political polarization that is pervasive in the U.S. Additionally, many platforms have since changed their policies, design, and business practices [141, 142], and many new platforms have been created and adopted by younger social media users since the onset of the pandemic [143, 144].

2.4.1 Young Adult Engagement with Politics Online

Young adults (aged 18-26 [145]) are predominant users of social media platforms, and similar to adults, they turn to social media platforms to learn more about political issues and

to share resources and information [118, 146, 49]. Additionally, young adult engagement with online politics and activism has been found to be indicative of their offline engagement with sociopolitical causes [147, 148, 149].

Prior research has explored young adult perceptions of their political engagement online and how it affects how they engage. Studies have found that although younger people turn to social media to inform their political opinions, they are wary of publicly sharing their opinions. A 2015 study with high school students in Sweden found that while high school students felt comfortable engaging in political discussions with their friends, they felt uncomfortable doing so on social media due to its public nature and how quickly conversations can escalate into arguments [150]. Similarly, a 2018 study with young adults in Hong Kong found that while participants felt comfortable with physically participating in protests, they were apprehensive about sharing their political opinions on their social media due to the argumentative nature of those spaces and feared engaging in online political discussion would negatively affect their friendships [18]. These studies and others have found that younger people are apprehensive about engaging in political discussion online because these spaces of discourse dissolve into spaces of argumentation [151, 152]. This leads them not only to avoid posting or commenting in these spaces but also to be largely skeptical of the effectiveness of social media being used as a tool for political discourse or political change [153, 154, 155].

Prior research has examined the effect of young adult political activism on social media, such as in the 2011 Arab Spring [52, 53], and the 2008 U.S. Presidential election [51, 156]. However, the young adult users of social media in much of this research are members of the Millennial generation (1981-1996) [47]. The context of social media use has changed significantly since 2011 in its scale and impact on public political life. For instance, in 2011, Facebook had approximately 750 million users [157], however, at present, Facebook has over 3 billion users globally [158]. Since 2011, social media platforms have significantly affected election outcomes globally [159, 160] and have played significant roles in

enabling the genocide of marginalized communities such as in the case of the genocide of the Rohingya people in Myanmar [161, 162]. In the United States specifically, social media platforms such as Facebook and Twitter have helped facilitate sociopolitical awareness campaigns and protests such as the Black Lives Matter movement [163, 17], but also have been pivotal in disseminating false information regarding the COVID-19 pandemic [164], and regarding election information as seen in the 2016 and 2020 U.S. Presidential elections [25, 54]. Terms such as 'fake news', misinformation, and 'filter bubbles' have entered the lexicon of social media users across the globe [55, 165].

For members of Gen Z, born between 1997 and 2012, social media platforms have been omnipresent throughout their lives. This constant, long-term exposure to the effects of social media in the context of the geopolitical sphere provides them with a unique insight into how to navigate online political spaces. For young adult women specifically, the offline effects of social media discourse have affected their perceptions of these spaces of discourse, in addition to shaping their behaviors within these spaces [68, 166]. While prior work in young adults online political participation has explored how they engage in these online spaces, much of this work has not explored the potential effects that gender identity has played on online political participation among young adults. Furthermore, much of this work has been done prior to the onset of the COVID-19 pandemic, with some exceptions [50, 167, 48]. As a result, much of the conceptualizing work about young adult online political participation does not fully reflect the effects of misinformation, polarization, and the turbulent political climate that affects social media usage and discourse in the current U.S. context. The social media political climate in which Gen Z young adult women engage drastically differs from that of young adults in the late 2000s and early 2010s, who are part of the Millenial (1981-1996) generation. The offline influence and impact of one's online behavior is more widespread and understood by Gen Z young adults [48]. Thus, I examine how this long-term exposure and knowledge of social media affect Gen Z young adult women and their perceptions and forms of engagement in social media political discourse

spaces.

2.4.2 Women and Political Engagement

Women's engagement in online political discourse spaces has been largely studied in the fields of political science, linguistics, sociology, and public policy, as well as HCI and social computing.

For many women, social media platforms have facilitated the creation of spaces in which they are able to openly share and connect with one another over personal and emotional experiences such as childbirth and sexual trauma [168, 169]. In some instances, online spaces have enabled women, who otherwise may not have felt comfortable or were unable to do so, to express their political stances [170]. However, while some have had positive experiences with the connectivity that social media platforms provide, many women have experienced hostility and harassment online, specifically in political spaces.

Prior research examining forms of engagement within news and political spaces online has found a prominent gender gap in public forms of engagement such as commenting and posting. Many of these studies have found that women prefer to read, lurk, and like other posts over commenting or making their own posts [59, 57, 58]. A study looking specifically at news-oriented subreddits found that although women users vote and read just as much as other users, they are less likely to comment or make a post [171]. Such results are indicative of a desire to engage with the content in these spaces, but a preference to avoid public forms of engagement within these spaces.

Other prior work in this space has found that women in these public discourse spaces often face gendered harassment and other identity-based attacks during political discussions online [62, 60, 172]. This online abuse often forces women to self-censor and avoid sharing their political opinions out of fear for their own and their family's safety [67, 27]. This has led to research developing new policies, practices, and systems for the content moderation of online abuse [173, 174, 175].

Nevertheless, this hostility and toxicity serve to broadly deter women from publicly participating in politics offline. Despite repeated calls for action for social media companies to improve their online abuse moderation practices, many women politicians have cited large amounts of unmitigated harassment and abuse as a reason for their departure from politics [176, 177, 27]. The fear of potentially violent consequences for sharing their political opinions online and or pursuing a career that is public-facing inhibits women from participating in politics, and as a result, pushes them to the periphery of these spaces, which significantly impacts the functioning of democratic processes.

The work of this dissertation focuses on the experiences of young adult Gen Z women in social media spaces of political discourse in the United States. However, in limiting the scope of this work, I do not address other intersectional characteristics and identities that affect engagement in online spaces, such as race, sexual identity, class, location, language, socioeconomic status, education, and many others. These dimensions of one's identity also shape one's relationship to spaces of online political discourse and one's perceptions of 'free speech'. I acknowledge that my focus on the broader experiences of young adult Gen Z women limits the understanding of how nuanced this experience may be when accounting for the various intersectional dimensions that affect this experience. The work of this dissertation serves as just one of many potential explorations into how social media practices and policies shape engagement in public forms of online political discourse among different populations of people. Extensive future research is needed to untangle the many intersectional dimensions that influence and shape one's experience online and to expose the tensions underlying the relationship between these dimensions and interpretations of 'free speech' in online political discourse spaces.

2.5 Next Chapter

This dissertation draws on this prior work to explore the influence of US-interpretations of free speech on social media practices and policies, specifically in the context of how young

adult Gen Z women engage with political discourse. If social media moderation practices and policies 'shape the shape of discourse' [16], what role does the underlying free speech theory play in that? Furthermore, how might these policies and their conception of 'free speech', shape how young adult Gen Z women engage in political discourse? In the next chapter, I explore Gen Z young adult women's perceptions of and engagement in social media spaces of political discourse. Specifically, I present the findings from an interview study and a diary study focused on understanding how Gen Z young adult women choose to engage in these spaces and what serves to encourage and or inhibit their engagement.

CHAPTER 3
FORMS OF ENGAGEMENT IN ONLINE POLITICAL DISCOURSE AMONG GEN Z YOUNG ADULT WOMEN

3.1 Introduction

Public and open political discourse is considered a vital part of the democratic process. Through discourse, ideas are debated and concerns are shared, and should these discourse spaces become inaccessible, democratic processes would be significantly negatively affected. As social media platforms have become more prevalent, research has shown that much public political discourse occurs online [118, 117]. These spaces for discourse are widely accessible and expose individuals to a wide variety of political ideologies [119, 178, 179]. Prior research on young adult (18-26 years) [145] engagement in online political spaces has focused on how social media platforms have been central to facilitating political discourse and how online political participation is indicative of offline political participation among young adults [50, 49].

Young adults use social media to perform activism, seek and share resources, and engage in political discussion [167]. Recent research has also explored how polarization and hostility have negatively affected young adult engagement in online political discourse [18, 150], and how online toxicity and harassment have negatively affected young adult women's broad engagement in online spaces, indicating gender-based differences in experiences when engaging in online political discourse [68, 152, 166].

In 2024, young adults between the ages of 18-27 were born between the years 1997 and 2005. Those in this demographic have been colloquially referred to as 'Generation (Gen) Z' (1997-2012) [47]. Prior HCI work has closely explored young adult online behaviors [180], however, unlike previous generations such as Millenials (1981-1996) and Generation

X (1965-1980), members of Gen Z have been exposed to the Internet and social media culture for a significant period of their lives [47, 48]. Many members of Gen Z were born after Facebook was founded (2004), and most members of Gen Z were born after Google was founded (1998) [181, 182, 183]. This long-term exposure provides them with a unique perspective on how the Internet and social media culture have shaped offline behaviors and culture [48, 184].

This chapter investigates: (RQ) How do Gen Z young adult women perceive and engage in social media spaces of political discourse?

The reason for focusing this work on young adult women in the United States specifically is rooted in their experiences and long-term exposure to social media culture. Prior work on women's engagement in political discussion spaces has shown that the growing hostility, polarization, and argumentative nature of these spaces have been shown to limit women's desire to engage in these spaces [185, 186, 187].

As Gen Z young adult women are amongst the newest group of voters in the U.S. and have been exposed to social media culture for an extensive period of their lives, I aim to understand whether their experiences with social media political discussion spaces negatively affect their perceptions and forms of engagement in political spaces more generally [48, 184].

Through a set of two studies, I closely examine Gen Z young adult women's perceptions of and engagement in online political discourse spaces, focusing on what serves to inhibit and or encourage their engagement in political discourse online, both publicly and among their selected social media community. Given the concerns regarding the withdrawal of women from public political participation in the United States due to the increase in unmitigated online toxicity [67, 27, 188], understanding how Gen Z young adult women, who are amongst the newest group of voters, perceive and engage in political discourse online is important when evaluating the efficacy of political participation spaces. As Gen Z young adult women, between the ages of 18-26 [145, 47], are of voting-age in the United States

(18 years) [189], I use the terms 'voting-age Gen Z women', 'young adult women', 'young adult Gen Z women' and 'Gen Z young adult women' interchangeably to describe this population of individuals in this chapter.

The first study discussed in this chapter is an interview study conducted with 17 voting-age Gen Z women across the United States (ages 18-26, born between 1997-2005) to examine their attitudes and perceptions towards political discourse spaces on social media, and how they choose to engage in them. Through these interviews, I explored their perspectives on what serves to encourage them to engage, what serves to inhibit them from engaging, and what strategies they employ when they engage in online political spaces. The second study discussed in this chapter is a two-week diary study with 13 young adult women (ages 18-27, born between 1997-2005) focused on understanding their motivations behind engaging and or disengaging from spaces of political discourse on a social media platform of their choosing. Of the 13 diary study participants, 7 also participated in the prior interview study. Their double participation will be noted and discussed throughout this chapter.

In both studies, I find that participants are critical of social media spaces of political discourse. They are skeptical of the political information they come across in these spaces, perceiving much of it to be false or misleading. They question the usefulness of sharing political information with their social media community since they feel that it does not lead to actionable change, and they largely find that social media is not conducive to engaging in productive political discussions, finding that most political discourse online becomes argumentative and toxic. However, I find that participants are not politically disinterested, as evidenced by their interest in keeping informed of political topics, and their preference for engaging in offline, in-person political discussions. I find that, as a result of their perceptions of online political spaces, participants deliberately choose to limit their political engagement online or to disengage from online political spaces entirely.

The work in this chapter contributes to research focused on understanding how social media platforms and culture impact political participation and democratic processes, by

connecting how perceptions of online political discourse spaces held by a group of voting-age Gen Z women lead to self-reported incidents of disengagement from these spaces. I draw on prior work on factors that affect young adults' and women's engagement in online political spaces to highlight how these participants' perceptions regarding the prevalence of hostility and argumentation in online political discourse spaces influence their choice to limit their presence in these spaces. Through this analysis, I 1) broadly challenge US-interpretations of free speech that are underlying the design and practices of social media platforms, 2) draw attention to definitions of 'political engagement' and 'political participation' in an online context that center on public forms of expression, to question whether a lack of public action and expression online equates to a lack of political engagement, and 3) question the efficacy of social media as a mechanism for public political discourse.

3.1.1 Political Participation and Engagement

There is much debate regarding the definition of political participation, and therefore no widespread consensus on a definition for online political participation [190, 191, 125]. Many definitions highlight 'public action' such as public expression, protests, and other such acts to influence political outcomes, as a key part of participation and engagement [192, 193]. Social media and online forms of engagement and discourse have also complicated interpretations and definitions of political participation. Prior work has examined and questioned whether 'slacktivism' and or digital activism should be considered part of political participation [124, 194].

In this chapter, I draw upon the work of Political Science and Communications researchers Kim and Hoewe to define and scope the terms 'political participation' and 'social media/online engagement' in the context of political spaces. Kim and Hoewe [190] identify five dimensions of **political participation**: 1) traditional political participation which includes donating, signing petitions, and protesting, 2) interpersonal political talk which includes having political discussions with friends and family, 3) voting in elections, 4) *so-*

cial media engagement, and 5) seeking information which includes reading and consuming politically-related content. Kim and Hoewe [190] further define **social media engagement** to encompass: 1) sharing an opinion on a social/political topic on social media, 2) expressing political opinions in online public spaces, 3) sharing political information posted on social media, 4) posting political messages online, 5) discussing politics online, and 6) 'liking' a social media post with political information. I draw on this definition of political participation because of its broad scope and its inclusion of online forms of engagement. I use Kim and Hoewe's conceptualization as a working definition but reflect on the limits of this definition in light of my findings in the Discussion (Section 3.5).

3.2 Methods

3.2.1 Participants (Both Studies)

All participants across both studies identified as Gen Z, born between 1997 and 2005, ranging from 18 to 27 (See Table 3.1, Table 3.2). I chose to focus on the experiences of voting-age Gen Z young adult women, as prior research on women's engagement in political discussion online has indicated that the growing hostility and polarization within these spaces has negatively shaped how women have engaged in them [187, 186, 185]. Furthermore, as Gen Z young adult women are amongst the newest voters in the United States, past work on how women are deterred from public social media political spaces [67, 57] indicates that this population of potential new voters in the U.S. specifically may be excluded from important political conversations that are occurring on platforms. Additionally, as aforementioned, Gen Z young adult women have had long-term exposure to social media culture, and during much of this time, social media has grown and changed, significantly impacting global political processes. Members of Gen Z have broadly observed the offline effects of online political participation and engagement [48, 195], potentially shaping this population's perspective on social media.

Participants were recruited through a combination of purposive sampling, snowball

sampling, and word-of-mouth. In order to get a geographically diverse set of participants, and avoid limiting the participant sample to Georgia, I used a number of different recruitment techniques. Recruitment messages were distributed among my personal networks, in three undergraduate courses at Georgia Tech, and on one university-wide mailing list for a research center focusing on women, science, and technology. I also requested participants to share the recruitment message with other potentially interested people in their personal networks. The selection criteria required that participants identify as women, be born between the years 1997 and 2005 (of voting age in the U.S. and a member of Gen Z), live in the United States, and use social media in some capacity 3-4 times a week.

Although potential participants were told that the interview would discuss their experience with political discourse, I did not ask about their level of interest in politics, or their political affiliation prior to recruitment.

3.2.2 Interview Study

Study Participants & Data

At the time of the interview, participants were located in different geographical locations which may have influenced their sociopolitical ideologies (however, I did not ask participants about their political ideologies as a part of the interviews). 3 participants lived in the Southern United States (Georgia), 1 participant lived in the Midwestern United States (Missouri), 1 participant lived in the Western United States (California), and 12 participants lived in the Northeastern United States (New Jersey, New York, Massachusetts, Maine, Vermont, Pennsylvania). Of the 17 participants, 9 identified as students from 6 different universities across the United States. The remaining 8 participants identified as currently employed or unemployed. All interviews were conducted virtually through video call over Zoom by the lead author. Interviews were between 60 to 70 minutes and were conducted between October 2023 and December 2023. The research protocol was approved by the IRB at Georgia Tech.

In the interviews, participants were asked about their broad experiences using social media and were free to talk in-depth about any platform of their choosing. The most discussed platform was Instagram, with TikTok and Snapchat being close seconds. I found that all participants were active on Instagram, even if it was not their primary social media platform of choice. I also observed that many participants would discuss the same online posts during interviews, indicating the virality and popularity of certain politically-oriented content on platforms such as Instagram and TikTok. I viewed these shared experiences of viral political content among participants as an indication of saturation.

Participants were asked about their understanding of political discourse and participation online, the types of political discourse they came across on social media, and if they chose to engage with it. If participants did choose to engage with political discourse, they were asked more specifically about the ways they chose to engage. If they did not engage with it, they were subsequently asked about why they chose not to. Participants were also asked how they would define political discourse and participation.

After the interviews, participants were asked to answer a set of four demographic questions to provide details regarding their race/ethnicity (as defined by them), their highest level of education completed, how long they've been an active user on social media, and how they would describe their location setting (e.g. rural, suburban, urban). Three participants chose not to answer these demographic questions (P7, P12, P15) (See Table 3.1, Table 3.2).

Data Analysis

I utilized an iterative coding approach when analyzing the interviews, informed by an inductive thematic analysis approach [196]. A broad set of themes emerged upon the first review of all interviews, using descriptive and process coding to identify topics and actions discussed by participants [197]. Following an initial round of coding, I wanted to understand how emergent themes aligned with existing conceptualizations of political en-

gagement, drawing on Kim and Hoewe's definition of political participation as mentioned in Section 3.1.1 [190], in addition to prior literature on online political engagement [18, 155, 151]. As such, I then took an abductive analysis approach [198].

As the focus of this research was on the connection between how the perceptions held by participants affected the way they chose to engage on social media, there were two broad categories that all codes were split between: 'Perceptions' and 'Engagement'. Codes that were under the 'Perceptions' category were, "Perceived Lack of Value of Online Discourse", "Argumentation", and "Perceived Lack of Knowledge". These codes focused on the perceptions participants held regarding online political discourse. Codes that were under the 'Engagement' category were, "Potential Online Consequences", "Potential Offline Consequences", "Anonymity", "Information Seeking/Sharing", "Lurking", "Disinterest", "Taking it offline", and "Avoidance". These codes focused on the ways participants engaged with spaces of political discourse, and potential factors that affected their online engagement in these spaces.

3.2.3 Diary Study

Following the interview study, I conducted a diary study between March and June 2024 with 13 young adult women to better understand their forms of engagement in online political spaces. 7 of the 13 participants had also participated in the previous interview study (P2, P3, P4, P6, P11, P13, P16). In this chapter, I denote participants who took part in both studies as (P #, Interview + Diary Study). Participants were asked to submit a diary entry twice a day reflecting on their engagement in political discourse on a social media platform of their choosing. They were asked to digitally send their diary entries to an anonymized researcher profile on the Signal app. Their entries were allowed to be text-based or audio-recorded. Following the study, I conducted a post-study interview with each participant in which they were asked about their experience in the diary study.

The decision to request participants to share their diary entries in a digital format was

to facilitate an easier transition between viewing a post on social media, reflecting on the post, and sharing the reflection. Furthermore, this decision was also informed by the age range of participants (Gen Z) and their comfort levels with using digital mechanisms to share their thoughts. I requested participants to submit two diary entries per day over the two-week period.

Each participant received a reminder twice a day through the Signal app to complete a diary entry. A reminder was sent once between 7-9 AM EST and then once again between 7-9 PM EST. In each reminder, participants were provided three guiding questions to write their entries. These questions were:

1. What is the political discourse you came across?
2. How did you feel seeing this post?
3. Why did you choose to engage or not engage with this post?

Although participants were encouraged to be informal in their reflections, the questions were shared as a starting point for the reflection process.

Following the two-week diary study period, I conducted a 45-minute post-study interview with each participant. The purpose of the post-study interview was to ask participants about how they felt about their forms of engagement in political discourse spaces and about their general experience with the entire study. The post-study interview was conducted over Zoom.

Study Participants & Data

The diary study participants were from various states across the United States. 1 participant was located in the West (California), 3 were located in the South (Georgia), and 9 were located across the Northeast (Maine, Vermont, New Hampshire, Massachusetts, Pennsylvania, Washington D.C., New Jersey). Of the participants, 4 identified as students at the time of study, 2 identified as recently graduated, and 7 identified as currently employed.

The most frequently used social media platform among participants was Instagram (See Table 3.1,Table 3.2).

Data Analysis

To analyze the diary entries, I utilized an inductive coding approach. As the purpose of this study was to encourage participants to reflect on their forms of engagement in online political discourse spaces, I split codes into two categories 'Perceptions' and 'Engagement', similar to the analysis process in the previous interview study. The codes from the previous interview study also served as a starting point when analyzing the diary entry data. I went through each participant's diary entries and applied the inductive coding method. Under 'Perceptions' the codes were, 'Desensitization', 'Skepticism', 'Feelings of Helplessness/Stress', and 'Focus on topics directly relevant to them'. Under 'Engagement' the codes were 'Engaging offline', 'Forming an opinion', 'Reading comments', 'No commenting', 'Reposting on Instagram story', and 'Private discourse'.

As aforementioned, following the two-week study period, I requested to conduct a post-study interview with participants. I similarly engaged in an inductive coding approach to analyze these interviews, drawing on the codes from the interview study and the diary entry analysis as starting points. Splitting codes into two categories, 'Perceptions' and 'Engagement', I found the codes under 'Perceptions' to be 'Performative Activism', 'Fear of negative social repercussions', 'Concern for safety', 'Feelings of stress', 'Shared context increases comfort', 'Awareness of ecosystem', 'Focus on topics directly relevant to them', and 'Skepticism of effectiveness of social media'. Under 'Engagement' the codes were, 'Engaging offline', 'Shallow engagement', 'No engagement', 'Forming opinions', 'Reading comments', and 'Creating posts'.

3.3 Perceptions of Political Discourse on Social Media

The findings both expand upon and complicate prior results that show how young adults may intentionally limit engagement in online political discourse. Consistent with prior research on young adults' engagement in online political discourse, I found that participants avoided publicly engaging in online political discourse spaces [18, 154, 167]. However, expanding upon past work, I found that participants were critical of social media as a space of political discourse. They were largely skeptical of the credibility of the political information they came across on social media. They questioned the purpose of sharing political information and resources on social media as they felt it did not lead to meaningful offline change. I also found that participants spoke of simultaneously feeling desensitized yet stressed while in online political discourse spaces. Most participants felt that social media was not conducive to discussing politics. These perceptions led to participants either limiting their online political engagement or disengaging from online political spaces entirely, which I describe in the following section (Section 3.3.1).

3.3.1 Skepticism of Online Information

Participants were largely skeptical of the credibility of the information shared on social media. P2 and P17 spoke of taking in information posted to Instagram and TikTok with 'a grain of salt', as they felt that most of the political information was partial. P15 shared this perspective, further describing her prior experience coming across false resources and information on Instagram during the 2020 Black Lives Matter Protests,

> "I feel like there was a lot of misinformation going around in 2020 with the Black Lives Matter protests. People just sharing something to share something. And I don't think a lot of them would check their sources." - (P15, Interview)

P15's statement also indicates a broader skepticism regarding the purpose behind one's politically oriented post and how it may affect the credibility of the information shared. P4,

P10, P14, and P17 similarly felt that many politically oriented forms of online engagement were done as a way to display one's sociopolitical values, but often resulted in the dissemination of misinformation, since the information was shared quickly without any verification.

Additionally, P11 felt that the prominence of advertising on platforms like Instagram led to users posting misleading information without verifying it, to benefit financially,

> "These content creators will post something about reproductive rights, they'll advertise something and then I'm like, you don't have like enough background that you're giving to your users to properly convey the message... Like, there's other sides of it and you're only giving one context because you're trying to push your agenda." - (P11, Interview + Diary Study)

Similar to P15, P11 felt that the purpose behind one's post could affect the credibility of the information shared. However, it is evident that P11 was further skeptical of the broader sociotechnical system in which the post was made, drawing attention to the capitalistic incentives behind platforms like Instagram driving the dissemination of false or misleading information.

P3 shared P11's concerns regarding capitalistic incentives on platforms such as Instagram, specifically in the context of charitable donations. In a diary entry she submitted, P3 wrote about an instance she came across on Instagram in which users may have been falsely advertising charitable donation links for citizens in Gaza. She wrote in her entry,

> "Today I saw an Instagram story post sharing links to donate for different forms of Gaza relief. The post just had Venmo accounts, which felt like they would go more directly to people in need, but at the same time are harder to trust. Again, this is the type of post that's difficult because it's shared by someone who I don't know well but trust on this topic, but I don't have the energy or know-how to verify these sources myself. While I would like to support relief efforts,

> again, I'm always unsure how significant my contribution will be, especially with the uncertainty about who it's really going to." (P3, Interview + Diary Entry)

Participants frequently questioned the source and credibility of the information they came across. It is evident that they have a knowledgeable, and largely pessimistic, view of the practices by which partial or misinformation is shared on social media.

3.3.2 Skepticism of Offline Political Change

Many participants felt that sharing political information among their social media communities on Instagram or TikTok was unproductive because it did not lead to effective offline change.

P4 described coming across infographics about the Israel-Hamas war, shared daily by someone they followed on Instagram,

> "Every single day, it's an infographic about Gaza and the Israeli apartheid, et cetera, stuff like that. And it's like, I agree with you, but what are you accomplishing by sharing that same information to the same group of like 113 people every single day?" - (P4, Interview + Diary Study)

P4's statement reveals that the audience to whom such information is shared is an important aspect of the skepticism participants have regarding the sharing of political information on social media. Nearly all participants followed and were followed by a selective group of users, many of whom they knew offline, and more importantly, were likely to share similar political views and ideologies. Thus, participants felt that sharing information with a wide group of followers who shared similar political ideologies was redundant and would not alter one's political stance. Diary study participants P21, P22, and P23 shared P4's perspective and reflected on this in their post-study interview. P23 noted that she, and many of her peers, deliberately obscure their names on Instagram to make them hard to identify

for privacy reasons. This then limits their reach and their network to only users they know in person, affecting the efficacy of trying to raise broad awareness through digital platforms in general.

> "I've been trying to think through this because I have some friends who make their Instagram account like a really obscure name or something that's not easy to find. . . they try to scrub a lot of information of themselves online.... But then it makes me wonder why we still feel compelled to post on our stories and everything if the main people who are seeing that are ourselves. Because it's like I still appreciate the post that I see my friends post , but if I'm receiving content about something that I've not necessarily been aware of before, I am more likely to think that the content is being posted for people like me who haven't been aware of it. That is like for the purpose of raising awareness. But I wonder at what point your account is limited in terms of who's aware of what, and it just becomes a little bubble. " (P23, Diary Study)

P22 felt more cynical regarding the sharing of political information on platforms such as Instagram as she felt that users often felt a specific peer pressure to post or contribute to the conversation despite their lack of interest and or knowledge.

> "I think there is pressure to post about these things and if you're not posting, it's like, well why aren't you posting? Do you not care or do you agree with the other side? But in terms of what is actually going on in the world, putting something on your story changes nothing. And if you really care about an issue, like actually do something about it or research it or talk to people offline. . . like do something about it if it's something you really care about versus like you put on your story— the story expires in 24 hours and you don't think about it again" (P22, Diary Study)

Her statement reflects an understanding of the performative aspects of social media,

specifically in the context of performative activism. P4, P10, P13, and P20 shared a similar perspective on users posting about political content out of peer pressure rather than a genuine interest in the topic they were posting about. Their perspectives also support the general skepticism participants had regarding meaningful offline change as a result of raising awareness or engaging in political discussions online.

These participant perspectives also suggest that the size of one's social media following affects the choice to engage with political content on the platform. P10 felt that political change was more likely when someone with a large following was discussing or raising awareness about an issue,

> "I think people give too much power to social media. Like if you're someone who has a lot of influence and power and you are posting links to charities and to rallies, like actually raising awareness of things that could be linked to some actionable change, then great, you should do that." - (P10, Interview)

P10's statement indicates that the size of one's social media following lends them some degree of 'power' and 'influence', suggesting that for those without an audience that gives them power and influence, posting political statements on social media seems useless due to their limited social media following and lack of widespread reach.

The potential results of one's engagement are important to participants in their decision to engage in online political discussion spaces. Participants felt skeptical of sharing political information on their social media because it would not change or influence their followers' political stance, nor would it be shared widely enough to enact significant political change.

3.3.3 Fear of Negative Social and Physical Repercussions

Social Repurcussions

Many participants were worried about potentially offending someone by sharing their political opinions and facing negative social repercussions as a result. For instance, P4 discussed their fear of being ostracized by their chosen Instagram community, many of whom they are friends with offline, if they posted something that was not in line with their community's values:

> "If I shared something that violated the understood values of my community...I would face social repercussions...people might ghost me. Like my friends might stop talking to me..." - (P4, Interview + Diary Study)

P17 discussed how she felt that she needed to scope out the topics or issues that she could 'broach' before engaging in a discussion with someone to ensure that the conversation remained respectful,

> "People tend to get really sensitive around the topic these days. So in order to keep away from ruining a friendship, I try to understand what topics I can broach and like what topics we can have a sort of respectful conversation about." - (P17, Interview)

P18 and P20 share a similar perspective to P17 in that they were both cautious about broaching specific topics with people they considered close friends. Both participants felt that their close friendships would be negatively affected if they engaged in political discourse.

> "I'm not voicing my opinions on the internet. Part of it is that I'm conflict-averse when it comes to things like this. Like one of my good friends I'm pretty sure is a Zionist. Like they are someone that I would reach out to if I

> needed help. Um, and they are moving to California, and I do intend to visit them. So like they're close enough where I feel like I'm comfortable enough to stay with them, but I'm not comfortable enough to like bridge this topic [Israel-Hamas war] with them." (P20, Diary Study)

From P4, P17, and P20's statements, it is evident that some participants felt that it required a lot of effort to identify people they felt comfortable speaking about their political opinions. The anxiety around potentially facing social repercussions by their selected community served to deter participants from sharing their political opinions openly online.

Participants feared their political stances might offend or hurt others and they would be judged by their peers as a result. The consequence of this was that participants largely avoided sharing their political opinions with their wide social media community.

It is evident that participants prioritized their personal relationships over publicly sharing their political opinions. This finding challenges pervasive Libertarian notions of 'speech' in the United States which interpret 'speech' as actions that are discrete and separate from the context in which the speech is made. I discuss this further in the Discussion section (Section 3.5).

Physical Repercussions

A few participants spoke of the fear of experiencing physical repercussions for sharing their political opinions publicly online. P19 prioritized her safety whenever engaging in political discourse, going so far as to agree with opinions she did not agree with in order to mitigate a hostile conversation.

> "If I don't feel safe, I'm like, you are right. Everything you say is completely fine." (P19, Diary Study)

P4 spoke of a fear of being physically attacked for speaking openly about her sexual identity in certain spaces. This fear of being targeted for one's identity was shared by P18

who spoke of their position as a religious minority. During her post-study interview, she spoke in-depth regarding her family's past experiences with political trauma, which shaped her own perceptions of how to engage in political discourse both online and offline.

Additionally, many participants spoke of a fear of facing professional repercussions for engaging in public political discourse. For instance, P5, P7, and P10 spoke of potential negative professional repercussions for publicly sharing their political opinions regarding the Israel-Hamas war. P5 spoke of colleagues getting fired for taking a strong public stance on the conflict. P7 spoke of friends being blacklisted from specific job opportunities for speaking out in support of Palestine.

The fear of facing social and or physical consequences for publicly aligning oneself with a political stance also surfaces the question of whether these young adult women participants are actually enabled to speak 'freely'. These experiences that participants shared illustrate an increasingly hostile, polarizing political environment within the United States regarding the conflict in Gaza that has, in effect, demonstrated to young adult women that engaging in 'free speech' may affect their ability to remain employed or physically safe. If they believe they will be negatively affected physically, socially, and or financially for speaking openly about their political stance, do they have the ability to engage in 'free speech' in the way free speech absolutists interpret it? I discuss this further in the Discussion (Section 3.5).

3.3.4 Feelings of Stress and Despair

Across interview and diary study participants, there was a common feeling of general stress and despair regarding the current state of politics in the United States. Many participants felt a sense of helplessness regarding their ability to make impactful change, and some spoke of feeling desensitized to events going on due to the onslaught of political content they were exposed to. P13 wrote in their diary entry,

> "The political discourse I came across was a Palestinian voter who is voting

> for Trump for the 2024 election. A friend reposted a post on her story. I felt weird because this year's election is weird. Like it sucks that we're stuck with the same 2 candidates as last time and it's also weird that people are voting for Trump because Biden was not good. Idk if I'm even voting this year but we'll see. . . '" (P13, Interview + Diary Study)

Multiple participants shared P13's ambivalence regarding the 2024 U.S. Presidential election. It is important to note that the interview and diary study were both conducted prior to Vice President Kamala Harris becoming the Democratic nominee for the 2024 election.

P16, P20, P21, and P23 used words such as 'depressing', 'hopeless', and 'doomsday' to describe their perspective on the current geopolitical situation. Through their diary entries and interviews, they reflected on their fears and anxiety surrounding politics and their lack of ability to create change. Some participants spoke of how their feelings of stress and anxiety led to them deliberately avoiding political content online. P6 reflected on her avoidance of all content relating to Israel and Gaza,

> "Like the Israel-Gaza war. Obviously I know these things are happening, but I never tried to learn more about it. And I feel like a horrible person saying that. But I never really dove deep into it to learn about what the issue is, who's involved, what are everyone's stances on it. . . it's not that I don't wanna know. I wanna be aware, but not to the point where I start to think about it a lot. 'cause I do think a lot. Like, I have to work. So I don't like to be thinking about, oh, the war is happening, what is happening to all these innocent people, like, while I'm working." (P6, Interview + Diary Study)

Many participants who spoke of avoiding these spaces also reflected on their privilege in being able to do so as someone living in the United States. Some reflected the continuous barrage of information provided to them on platforms such as Instagram and Twitter

resulted in a broad sense of desensitization. P2 wrote about this in a diary entry she submitted,

> "Today, to be completely honest, I had a fairly unplugged day. I did not come across anything that stood out as political discourse on social media today. I felt relieved in a sense that I didn't have to engage with such posts today. While I believe that political discourse is of the utmost importance on social media, I think being constantly bombarded with different posts, infographics, and graphics can be tiring and make one more normalized to seeing such posts, thereby becoming desensitized to posts in the future." (P2, Interview Study + Diary Study)

3.3.5 Critiques of Platform Practices and Affordances

Through these interviews and the diary study, I found that nearly all participants felt that online spaces of political discourse were argumentative and hostile, consistent with prior work [130, 151], which negatively affected the productivity of discourse that occurred in these spaces. However, I found that they also had notions of what features of social media platforms contributed to the hostility found in spaces of political discourse, and why these features made social media platforms not conducive to political discussion.

Critiques of Short-Form Content & Quick Engagement

Some participants specifically discussed how they felt social media spaces were not conducive to discussing politics due to platforms' practices of encouraging the creation of short-form content that enables quick engagement. P3 described how short-form content reduced the complexity of a political issue,

> "I feel like there are limits to what you can say on social media. It's meant to be more of this, you know, small bits of little information, easy to consume

quickly. So I think especially with complex issues...I feel like it's harder to really explain anything on a social media platform or even online generally..." - (P3, Interview + Diary Study)

Similarly, P11 highlighted how the emphasis on short-form content on platforms like Instagram negatively affects how people interpret the political information they see online because short-form content is made to draw clicks and attention and does not allow for a full exploration of complex issues,

"I think like any big news event you see a lot of really heavy sides when you're on social media. And because it's short-form content, you don't see the full image. It's like the focus is again, how many clicks can I get?" - (P11, Interview + Diary Study)

Other participants supported this, further discussing how short-form content may lead to argumentation, and or hostility due to the potential for misreading and or misunderstanding. P9 and P15 both described how the fast engagement on X (formerly Twitter) led to argumentation.

The discussion of short-form content and how platforms encourage quick engagement reveal that participants were reflective on the potential reasoning for why they were not comfortable engaging in social media spaces of discourse. It also further demonstrates their awareness of social media practices and their attention to how they and other users around them are being affected by these practices.

Concerns about Algorithmic Feeds

Beyond concerns about the length and type of content, several participants expressed concerns about how political engagement online might affect the type of content that is shown to them through the platforms' algorithmic feeds. P2 felt that liking and reposting political information that was shared among her selected Instagram community led to Instagram's algorithm prioritizing information that supported her views,

> "The algorithm prioritizes only what you want to interact with. So I think that sometimes it can turn into a bit of an echo chamber, where if you're only interacting positively with certain types of posts, you're only going to see that type of view reinforced on your feed. And any type of dissenting views wouldn't necessarily be represented." - (P2, Interview + Diary Study)

P5 described her experience coming across highly misogynistic sociopolitical content on Instagram despite deliberately avoiding political content on the platform. She felt concerned that Instagram's algorithm was pushing forward such content and giving it further exposure,

> "I think it was a really popular video and then Instagram gave it to me. It's not that I'm searching out those videos...I don't know how they let people like that get big. I mean... make sure your algorithm doesn't promote people like that." - (P5, Interview)

While prior work shows how users engage in practices to try to exert some control over their algorithmic feed [199, 200, 201], these findings suggest that participants also intentionally disengage from certain actions (such as liking and reposting political information) as a way to exert control over algorithmic content feeds.

Critiques of Pseudonymity and Anonymity

Many participants also specifically mentioned that they felt not linking online identities on social media to offline identities (through platforms that allow for pseudonymous and anonymous accounts) potentially led to forms of aggressiveness and argumentation in online political discourse, which made them feel uncomfortable and stressed. P15, P16, and P17 all felt that since there was a minimal chance of meeting offline, anonymity enabled users to be aggressive without consequence. P17 further went on to describe how

anonymity enabled users to spread false information without consequence due to the lack of profile information,

> "Anonymity is very helpful for them because they can just talk and there's no information relaying them back to any of their social platforms or even any other social platforms... It's just like they can say it and no one's gonna say, Hey, you should fact check. Or they could say you can fact check, but they're never gonna know if they followed through or not." - (P17, Interview)

P9 concurred with the feeling that anonymity enabled argumentation, but also felt that since conversations on platforms like X and Reddit were semi-public, this facilitated political discussion to devolve into argumentation quickly since more people could be involved,

> "It always like gets worse and worse because, in an online setting, more people can get involved. It's not just you and another person. And depending on where you are on the internet, people can gang up on you. It's not just you and that person, it's other people chiming in and maybe they're a little more direct, so then you get attacked by this other person." - (P9, Interview)

Participants' awareness of the specific affordances on social media that affect their perceptions and engagement in online political discourse indicates that they hold an in-depth understanding of platform design and culture, in addition to their understanding of their role as users in this broader sociotechnical ecosystem. Participants not only identified that online political discourse environments were argumentative and hostile but also reasoned through potential platform features such as short-form content, algorithmic recommendations, and anonymity that influenced this argumentation and hostility.

Despite participants finding social media not conducive to political discourse, they found alternative ways to engage in political discussions. Below, I discuss how they engaged and disengaged.

3.4 Practices of Engagement

The perceptions of social media spaces of political discourse led participants to limit their online engagement or disengage from online political spaces entirely. I describe these practices below.

3.4.1 Limiting Online Political Engagement

I found that participants engaged in specific actions that limited their online engagement in social media spaces of political discourse.

Not Information Seeking or Sharing

I found that many participants did not seek out political information on social media. P5 described how, despite her avoidance of politically oriented posts on Instagram, she believes that the platform algorithm keeps showing her such content because it is popular. Furthermore, a few participants spoke about deliberately avoiding any news-related content on social media as they felt it was either false or misleading, preferring to read news from sources such as the *New York Times* or the *Wall Street Journal*. Additionally, I found that many participants did not share much political information, such as in the form of infographics, on their social media accounts, believing that the practice was ineffective in creating actionable, offline change.

Avoiding Public Forms of Engagement

Many participants avoided public forms of engagement, such as commenting. Many chose not to comment on public, politically relevant posts because they felt it was pointless since the comment would be lost among the potentially hundreds of others and would not be read. P11 preferred blocking users who made posts that she felt were problematic or disagreed with rather than commenting because she felt that the user would not care to read her

comment,

> "I just don't think people would read it. Like I've never been like, I can have this good response to a video that people would care about...I've thought about sending a DM once or twice where someone will post something and I'm like, oh my gosh, that's horrible. But then I just feel like maybe they won't see it and maybe my wording won't be good enough. So I'm like, whatever, I just block them and continue on. " - (P11, Interview + Diary Study)

While most participants deliberately avoided any form of public discussion or conversation, a few participants did engage and described their negative experiences doing so. P9 spoke about an experience she had engaging with an unknown user on X (formerly Twitter) regarding the efficacy of the COVID-19 vaccine. Following this experience, she felt that the conversation wasted her time and made her more angry regarding the situation rather than resolving anything,

> "It never goes well. It's always like they retort back and then maybe I'll retort back or maybe I'll just be like, I don't wanna respond anymore because it's obviously like wasting both our times. And then after I just reread it I'm like, oh my goodness. Like this is such, I don't know, it was a very annoying situation. So I always feel bad before, during, and after I engage in something like that. " - (P9, Interview)

P21 similarly spoke of an experience she had engaging in a discussion with an unknown user regarding abortion rights. During the discussion, the unknown user went through her social media profile and began using some of the images against her during the discussion,

> "But what they did was they stalked through my profile and found this one photo of me with a special needs student and used that against me in a way to like say I was wrong. And after that I was like, this is too much. So I just don't

> respond because I feel like it's not productive to be expressing my opinions in the comments section." (P21, Diary Study)

P1, P5, P6, P7, P13, P14, P15, P16, P17, P18, P19, P20, and P21 specifically mentioned that the hostility of these spaces led them to avoid publicly engaging in political discourse altogether. P9 and P21 explicitly spoke about the hostility they experienced firsthand, which then shaped their subsequent experiences and limited engagement in these online spaces.

These findings suggest that perceptions regarding social media spaces of political discourse have led participants to limit their outward engagement in these spaces, evidenced by their preference to read and consume political information rather than comment, post, or widely share their political stance.

However, it is important to note that despite limited public forms of engagement, all participants held political opinions based on the information they read and consumed. These participants were not politically disinterested, they were disinterested in publicly sharing their political stance online.

3.4.2 Creating Private Political Discussion Spaces

Nearly all participants mentioned that they were cautious with whom they discussed politics. Instead, most participants were selective with the smaller sub-group of people they discussed politics with. Individuals who were part of these private political discussions were often friends, family, or classmates whom participants felt shared a similar background or context. While they did not necessarily need to know or have met the individual offline, nor need to share similar political opinions to them, participants felt that they had to understand the context from which the individual's perspective was coming and had to feel comfortable that they would not offend them if they disagreed.

Many participants mentioned that conversations felt more consequential when they knew the identity of the person they were talking to. P17 felt that having a political discussion with someone she did not know made her question their perspective because she did

not know any information about their reasoning or background,

> "I think it's because there's no backing to it. There's no reasoning behind these figures and like who are they?... just no information on their profile at all. So you can't really gauge who these people are, what they do in their life, how affected they are by this. They can talk all they want, but there's no reason to believe them..." - (P17, Interview)

P17's statements reveal that she feels that credibility is directly tied to one's identity in social media political discourse spaces. Her statements indicate that knowing why someone is engaging in a discussion and how the topic of discussion is relevant to them lends them authority and credibility in the discussion.

P3 felt that when she discussed politics with someone she knew, she felt that she understood their perspective more because she understood them as a person, which in turn lent itself to an overall more productive conversation,

> "I would prefer to have those conversations where you can have more of a back and forth with someone that you trust and you value their opinion and you know that they're coming into the conversation wanting to learn something or like being open to views. Even if they do have a very specific view that's opposite to mine, if I know them as a person I think I can more easily be able to have that conversation with them..." - (P3, Interview + Diary Study)

P3's experience indicates that trusting a person, valuing their political opinion, and knowing the intention with which they are coming to the discussion are all factors that she feels influence whether the conversation will be productive.

3.4.3 Sharing Context

Other participants mentioned that even if they did not know the person they were talking to personally, sharing some degree of context with them made them more comfortable

talking to them. P11 discussed an instance in which she publicly commented on a post another student at her university made about protestors during the 2020 Black Lives Matter protests. Although she did not know the student personally, she felt comfortable engaging in a discussion with them because they both went to the same university. P16 felt that engaging in political discourse, online or offline, with others of a similar age was preferable because she felt they might share a similar perspective. P6 discussed this further, explaining that she felt some degree of comfort engaging in political discourse with someone more similar to her in race/ethnicity and gender identity, as she felt that they may share similar perspectives,

> "I will say though, if it was someone I didn't know, I would be more comfortable talking to a woman than I would a man.... if it's an Indian woman, even more so. I guess someone who's more similar to you. You're attracted to the people who are similar to you, right? Like, you become friends with the people who are similar to you, have similar interests to you. Now that I'm thinking about it, I think if they are similar to you in more than one way, you start to think that they also think similarly to you." (P6 - Interview Study + Diary Study)

It is evident from the experiences above that what it means to 'know' someone in the context of feeling comfortable engaging in productive online political discourse differs among participants. For P17, knowing the context of one's perspective on a political topic is sufficient to engage in political discourse. For P3, trusting the person, valuing their opinion, and knowing their intention behind entering a political discussion is important to engage in productive political discourse. For P11, knowing that they had a shared background or context, such as the same academic context, with whom they were talking enabled productive political discourse.

3.4.4 Expanding Offline Political Discussion

As an alternative to engaging in political discourse online, many mentioned preferring to engage in discourse offline. P15 mentioned that the face-to-face element of in-person discussion enabled more empathy as she was able to see all the nonverbal aspects of communication that she felt were lacking in online discourse,

> "If you're in-person, having a conversation I think in general is more fruitful than social media. You can get a lot more of the nuances out or I just think in general talking face-to-face is better 'cause you get the facial expressions, the nonverbal forms of communication, which you don't get over social media. "
> - (P15, Interview)

P17 similarly preferred engaging in offline discourse as she felt that face-to-face discussion led to fewer misinterpretations compared to online discussion. P3 and P6 found that engaging in political discourse offline in a small group enabled a back-and-forth between people, allowing more time to formulate responses and take in the perspectives shared. P18 shared this perspective and further went on to explain how in-person conversations felt more fluid while online conversations felt more permanent because there was a written record. In-person, offline conversations do not have such a permanent record that tracks one's stance on a subject, enabling them to change or adjust their stance, which enabled a less rigid, more nuanced discussion,

> "Online it's like, if you put it out there, it's permanently there, right? So even if you change your opinions you can't go back on things that you've said. But in person, because generally most of the time I'm having these conversations with people I know, I can go back two days later and say, Hey, I found this and I've changed my perspective and then I could have another healthy discussion about it if I want to. " (P18 - Diary Study)

Participants' responses and preference for offline political discussion indicate that they value the qualities that verbal and face-to-face interactions enable. Qualities such as non-verbal forms of communication that can be observed, time, and unlimited 'space' to explain a viewpoint and understand others' viewpoints.

3.5 Discussion

Through exploring the perceptions these participants held regarding social media spaces of political discourse, I found that they were largely critical of these spaces, which influenced their decision to either limit their online political engagement or disengage from these online spaces entirely. Participants' perceptions of social media spaces for political discourse and their subsequent practices of disengagement encourage us to re-examine whether platforms are enabling their users to engage in 'free speech' and what it means to be 'politically engaged' in an online context. The findings demonstrate that participants were highly knowledgeable of social media functionality and culture and how it may broadly affect their own behavior on the platform as a user. These findings also indicated a sense of resignation and acceptance of the potential harms or negative effects of present-day social media functionality and culture, evidenced by participants choosing to limit their presence in online political spaces or simply leave them. The findings also draw attention to the potential for voters with similar views to the participants of these studies to completely withdraw from political spaces entirely. I discuss these further below.

3.5.1 Limiting Political Engagement

The findings demonstrate that participants, while not choosing to completely withdraw from online political spaces, limited their presence, time, and exposure to them. These practices are a form of disengaging from these spaces. Prior research has explored how politically active young adults engage or disengage on social media, however, these works were focused on the experiences of activists and political organizers and did not focus on

gender identity as a dimension that potentially affected political engagement [18, 180].

The Gen Z young adult women participants reported using social media as a tool to stay connected with their friends, and to explore their hobbies and interests, rather than for online political participation. While participants preferred limiting their engagement or disengaging from online political discourse spaces, it is possible that these practices may exacerbate the creation of echo chambers and further polarization. Many participants were self-aware of their tendency to create their own echo chambers both online and offline. While some participants did seek out alternative viewpoints, most felt comfortable only engaging in discourse with those of a similar viewpoint. Prior literature on echo chambers in online political participation has examined how one's preference to talk to other people who hold similar political opinions impedes them from challenging and questioning their perspectives [119, 151, 202].

For many participants, their skepticism regarding the credibility and usefulness of social media political discourse spaces led them to limit their online political engagement and to create private political discussion spaces. However, continuing to engage politically in these ways may result in a cycle that exacerbates political polarization. Many participants indicated a growing aversion towards politics, with eight participants specifically expressing a distaste for politics generally due to its polarizing nature. The continuous engagement in such practices may lead to participants withdrawing from political participation offline as well which could have a significant negative effect on democratic processes.

3.5.2 Implications for Political Participation

While the participants of both studies demonstrated an interest in being politically informed despite limiting their online political engagement and or disengaging from online political spaces, it is not unreasonable to speculate that continuously choosing to disengage from online spaces of public political discourse could lead to a broader withdrawal from offline public political spaces. The participants' skepticism regarding the political information

they came across online and of the platforms themselves indicates a resignation that they must continuously question the validity and credibility of what they come across on social media. Furthermore, their fear of experiencing negative social and physical repercussions for publicly associating with a political stance and their preference for offline forms of political discussion indicates a desire to avoid public expressions and action. Prior research examining the effects of online harassment and abuse targeting women has shown that the process of withdrawal can be incremental, with women initially choosing to adjust their online behavior, limiting their presence in online spaces, to then later leaving the space entirely [63, 67, 46]. Prior research examining the effects of abusive behavior on women's political engagement has shown that this behavior has discouraged women from running for political office, in addition to postulating that the continued existence of this behavior could lead to women's broader withdrawal from public political spaces which would significantly affect democratic processes [27, 20]. While this research did not focus on the impact that harassment and abuse had on voting-age Gen Z women's online political engagement, one can draw parallels between the findings of this work and the findings of prior research examining the relationship between online abuse and women's political engagement.

The participants' general aversion to online political spaces may not necessarily be rooted in prior direct experiences of harassment or abuse but may be rooted in discomfort with the argumentative and hostile nature of these spaces and the fear that they may escalate to harassment or abuse. Their discomfort with the argumentative and hostile nature of these spaces has influenced their decision to disengage and or limit their forms of engagement in online political spaces, leading to concerns regarding whether this may influence their offline political engagement. Additionally, participants' choice to be selective with whom they engage in political discussion similarly reflects a desire to avoid hostility by choosing someone they trust and or who will not be offended by their perspective. Furthermore, participants' preference for offline political discussions indicates that they find face-to-face interactions facilitate more empathy and less hostility toward the person they are conversing

with. The findings indicate that while participants attributed their forms of engagement to their criticisms of social media spaces of political discourse, their discomfort regarding the hostility and argumentative nature of these spaces—and its potential to escalate into harassment and abuse—was a key factor that influenced their decision to limit their online political engagement.

3.5.3 Accepting the Limitations of Social Media Discourse & Gender Identity

Despite participants' criticisms of social media as a space of political discourse, many felt that there were no other spaces of discourse that were as pervasive as social media. They felt that despite platform limitations in enabling inclusive and productive political discourse, people would engage in political discourse on them regardless since social media platforms have a wide reach. Participants were resigned to this and adjusted their behavior accordingly, choosing to limit their online political engagement or choosing to disengage from the online political discussion space entirely. For many who accepted the limitations of online discourse, engaging in offline discourse served as the better option.

This finding is consistent with prior research on women's online behavior. Prior research has shown that after long periods of exposure to hostility or negativity online, women may experience resignation and acceptance and are likely to adjust their behavior online to accommodate or cope with this [166, 68, 46]. For instance, a 2018 study examining coping mechanisms women in academia employ when faced with online abuse and harassment found that women engage in 'anticipatory self-protection'. To minimize potential harassment, women in academia may refrain from discussing personal aspects of their lives and may avoid writing or talking about controversial subject matter [203]. These participants' perspectives suggest that even without direct experiences of abuse and harassment, Gen Z young adult women may adjust their behavior in response to perceptions of hostile and negative online spaces.

While participants did not explicitly discuss their gender as an influence on their per-

ceptions of social media political discourse and their chosen forms of engagement, it is probable that their gender identity influenced them to an extent. Their fear of experiencing negative social repercussions may have led them to limit the people with whom they discussed politics, indicating that participants adjusted their behavior in anticipation of potential backlash, something that has been found in prior feminist sociolinguistic work [63, 46]. Additionally, participants' decision to minimize their public forms of engagement, such as commenting and posting in spaces of political discourse, is consistent with prior work examining prominent gender gaps in public forms of engagement. Many of these studies have found that women prefer to read [58] and like other posts over commenting or making their own posts [185, 57, 59]. However, participants' resignation regarding the fact that social media was the predominant space to engage in political discourse and their subsequent decision to limit their engagement in it has significant implications for Gen Z young adult women's political participation more broadly.

Although many women online have sought out alternative forms of engagement to avoid hostility and toxicity and developed mechanisms to cope with the limitations of social media discourse, I raise the question, why should they have to? Why should a population of social media users feel the need to create ways to protect themselves online? Why do they not expect the platform to protect them? Furthermore, if these users are coping with the limitations of social media discourse by avoiding them and choosing to shift offline, it raises the question of whether they are enabled to exercise 'free speech'?

3.5.4 Not 'Free' Speech

The findings from these studies challenge assumptions about whether these platforms enable 'free speech' for all users. The right to 'Free Speech' in the United States is considered a cornerstone of the democratic process and enables individuals to express their opinions openly without fear of the U.S. government infringing on this right. Although the law protects against governmental interference, the term and concept of 'free speech' is pervasive

in alternative contexts in the United States. Various public spaces where one can engage in discourse, such as schools and universities, workplaces, and social media platforms are places where questions of 'free speech' arise and are debated [16, 76]. On social media platforms, questions of 'free speech' and censorship have been raised and debated since their inception [77], although private companies in the United States do not need to abide by free speech law. Despite this, almost all social media platforms promise a commitment to enabling 'free speech' and 'free expression' for users on the platform. For instance, Meta CEO Mark Zuckerberg defended the protection of misinformation and toxic political rhetoric due to a broader commitment to 'free speech' [79], and X owner Elon Musk called the platform a 'digital town square', "where a wide range of beliefs can be debated in a healthy manner without resorting to violence" [204]. However, the experiences of the young adult women who participated in both studies indicate that they do not freely express themselves on these platforms, surfacing questions regarding the efficacy of these platforms in enabling 'free speech'.

Participants partially attributed their decision to disengage or limit their engagement in public spaces of political discourse to fear of potentially experiencing negative social and physical repercussions for expressing their political opinions online (Section 3.3.3). While many participants spoke of facing repercussions in their personal lives, such as impacting their friendships and relationships, many spoke of fear for their safety or of a fear that their statements could impact their professional opportunities (Section 3.3.3). If one chooses to self-censor on a platform out of fear of facing repercussions that could affect them physically or financially, are they being enabled the ability to engage in free speech? When one must decide between retaining employment or expressing a political view, it is difficult to argue that they have 'free speech'.

One may argue that the online spaces that the participants were in did not facilitate conditions under which they could exercise their ability to engage in free speech. For instance, the pervasiveness of hateful speech, online toxicity that contributes to argumentation and

hostility, and misinformation were factors in their decision to disengage and or self-censor (Section 3.4.1). The findings from these studies indicate that perhaps these users are provided a nominal 'ability to engage in free speech', as in practice they face physical, emotional, or financial consequences for exercising free speech in political spaces online.

The participants' decision to limit their engagement in political discourse spaces largely goes against what social media platforms claim they enable for their users. However, platform practices are not solely to blame for the participants' practices of disengagement. Other factors, such as the increasing polarization and hostility surrounding politics in the United States and the hiring practices of companies all contribute to their decision to disengage. In the United States, the increasing polarization and hostility surrounding politics, both online and offline, in the past decade has resulted in various instances of offline violence in schools [205], workplaces [206], and toward public figures [67, 26]. Furthermore, one's online actions have been shown to affect their offline life, as seen in how workplaces have begun monitoring employees' or potential employees' social media profiles for controversial or highly politicized posts [207]. As such, those who are currently employed often censor their behavior online or add disclaimers to posts with politically oriented content.

The confluence of social media practices, the current geopolitical climate, and other social and cultural practices affect and influence the online behaviors of the Gen Z young adult women who participated in these studies, specifically in how they choose to disengage from expressing their political stances and participating in online spaces of discourse.

3.5.5 De-Centering the 'Public' in Political Engagement

The experiences of participants in these studies also lead me to question existing definitions of 'political engagement', as many of these definitions of 'political participation' and 'social media engagement' such as those of Kim and Hoewe [190], center public expressions of political opinions in their interpretation [192, 193]. Kim and Hoewe specifically

draw attention to public actions when defining 'social media engagement' in the context of political participation, such as posting and commenting on politically oriented posts. Such interpretations assume that without public expression of opinion and information, through activism and protest or verbal expressions in online public spaces of discourse, political participation does not occur. These definitions are strengthened by the pervasive understanding that social media platforms are centers of public political discourse [118, 208] and have played a significant role in many global political activities and processes [53, 159, 17]. This emphasis on the public expression of opinion in online political spaces has also previously been questioned by instances of hashtag activism and or slacktivism in which a social media user may publicly associate oneself with a sociopolitical cause or show their support for a cause through 'likes', making posts using a specific hashtag, 'reposts', and or changing their profile photo as a form of political participation [209]. While some prior work has questioned the efficacy of slacktivism and hashtag activism [209, 126], others have indicated such online acts may be important in publicly showing solidarity or indicating offline action [210, 194]. However, this raises the question, with the predominance of social media as a mechanism to facilitate political engagement, does a lack of any public social media engagement in political spaces equate to a lack of political engagement? For instance, does a reluctance to publicly associate oneself with a political stance on Instagram equate to political disinterest or political apathy?

The participants' forms of engagement on social media led me to question widespread understandings and interpretations of what it means to be 'politically engaged', specifically within an online context. I find that participants largely limited their forms of 'social media engagement' as defined by Kim and Hoewe [190], only turning to social media political spaces to periodically seek out political information, to stay informed and read political information, and to 'like' political posts that were of interest or relevance to them. These limited forms of 'social media engagement' did not include any public expression of a political stance, something many participants spoke of deliberately avoiding. However, the

findings also demonstrate that participants were not politically disinterested individuals. They actively consumed political information and many explicitly mentioned that they had thoughts and opinions on what they came across, but would prefer to discuss their thoughts offline with small groups of people they trusted, rather than in public spaces of discourse. These findings encourage me to reconsider whether public forms of expression should be central to a definition of 'political engagement'. The participants of this study were politically engaged and thought critically about their sociopolitical environment. They sought to learn about issues, and then discuss this information with others that they trusted would not judge them for their opinion or potential lack of knowledge. While their mode of engaging with the political lacked the public nature of how US society defines 'political participation', the participants of this work were certainly not politically disinterested or apathetic. The experiences of the Gen Z young adult women participants encourage us to reconsider the emphasis on the 'public' in widespread interpretations of 'political engagement'.

3.5.6 Challenges to the US-Interpretation of Speech

The findings from both studies indicate that the fear of potentially facing physical, social, and financial repercussions largely shaped participants' decisions to not publicly share their political opinions on social media. For instance, it was evident that many participants prioritized maintaining their relationships over publicly sharing their political opinions, as seen in P20's hesitance to broach the Israel-Hamas war with a friend she believed was Zionist (Section 3.3.3), or P4's worry over posting something that potentially violated the norms of her chosen community (Section 3.3.3). This decision broadly challenges the US-interpretation of speech that views speech as separate from the context in which it was made.

The findings of these studies challenge the notion that speech consists of acts that are discrete and separate from their contextual setting. Participants of these studies viewed their speech online as actions that could have great impact on their lives and relationships.

Who they were talking to, and in what context that conversation occurred influenced their decision to engage in political discourse or not (as seen in Sections 3.3.3, 3.3.4).

Feminist language philosopher Jennifer Hornsby discusses this in her work 'Free Speech and Hate Speech' when she argues that Libertarian interpretations of free speech are flawed as their interpretations of 'speech' are limited [211]. She argues that since such perceptions hold that speech is separate from the context in which it was said, these perceptions advocate for the protection of 'hate speech' or 'offensive speech' since they are 'just words'. The Libertarian interpretation of speech does not consider the context, background, and situatedness of speech. Recalling the SCOTUS decision in the 2011 *Snyder v. Phelps* case [85] (Seen in Chapter 2), the Court Opinion protected the hateful and harmful speech, stating that despite speech's powerful nature and its potential effects, the law of free speech would not punish the speaker. In that case, despite the speech targeting a grieving family and the emotional and mental stress it caused, the speech was considered individually by the Court, separate from the context it was made in. The US-interpretation of free speech views speech as 'just words' and often not as acts or in relation to the situated context.

The findings from both studies challenge this pervasive notion of speech, indicating that the young adult women participants of this work interpret speech as embedded or fundamental to their relationships, as it holds the power to negatively affect these relationships. They viewed speech as not separate from the context in which the speech was said and their relationship to whom the speech was said, questioning the US-interpretation of speech. This posits the question of whether the notions of speech and free speech that are predominant on social media platforms are shared by their users. The Gen Z young adult women participants who regularly engaged on platforms such as Instagram evidently held an alternative view of speech to the platform. Furthermore, their fear of facing physical or financial consequences for stating their political opinions also indicates that perhaps they question their own ability to express themselves freely on these platforms.

3.6 Limitations

This work explores perceptions that a group of voting-age Gen Z women holds towards online spaces of political discourse and how they choose to engage with these spaces. However, due to the limited sample size, it is difficult to understand whether these perceptions are widespread among Gen Z young adult women.

The selection criteria for both studies were intentionally broad. I sought to interview individuals with varying degrees of interest and participation in online political spaces, and as such, I did not ask participants about their online political engagement prior to recruiting them. Participants were not recruited based on their degree of interest in U.S. politics but rather were recruited based on their age range, gender identity, and social media usage. In both the interview and diary study, I reached a saturation point, as all participants shared similar experiences with viral politically oriented content that they saw on Instagram. Future work could explicitly recruit participants active on social media but who do not use Instagram to explore a broader set of experiences with online political content.

I interviewed individuals who identified as women and were between the ages of 18 to 26 at the time of study, which placed them in the Gen Z (1997-2012) demographic [47]. While I asked participants about their race/ethnicity and tenure on social media, I did not ask about other demographics that may impact and influence forms of engagement on social media, such as sexual identity and political affiliation. The reason for this was twofold. One was to limit the scope of the interview and diary study to closely examine one dimension of identity, gender, and its potential effect on engagement within online spaces of political discourse. The second reason was that I did not want to compromise the anonymity of participants, many of whom during both studies spoke about their hesitation and caution as to what they reveal publicly. For this reason, I did not ask for more details regarding their identity.

Additionally, I acknowledge the lack of diversity in race/ethnicity, location setting, and

educational background among participants. I do not make any claims that this participant sample is a representative sample of Gen Z young adult women in the U.S. However, I did not seek to examine these dimensions of identity and their influence on Gen Z young adult engagement in social media political spaces and therefore did not recruit for them specifically. I report on these factors for transparency purposes, but I do not focus on them for the analysis. In future work, I encourage and call for the exploration of other dimensions of identity and their influence and impacts on Gen Z young adult engagement in online political spaces.

3.7 Conclusion

In this chapter, I explored how a group of 23 voting-age Gen Z young adult women in the United States perceive social media spaces of political discourse and how their perceptions may influence their forms of engagement in these spaces. Through 17 semi-structured interviews and a diary study with 13 voting-age young adult Gen Z women, I found that participants were critical of social media spaces of political discourse and chose to limit their online political engagement, create private political discussion spaces, and in some cases, disengage from these online spaces entirely. Through these studies, I found that participants had a deep understanding of social media functionality and culture and how it affected their online behavior. Drawing on these findings, I 1) examine the efficacy of social media as a space for public political discourse, 2) question what it means to be 'politically engaged', arguing that a lack of public forms of engagement does not equate to political apathy, and 3) challenge widespread notions of 'speech' and 'free speech' by contrasting participants' interpretations of 'speech' and the pervasive US-interpretation. The consequences of Gen Z young adult women choosing to disengage from public political spaces are incredibly significant for democratic processes globally. Who speaks, who silences, and who is silenced are not questions of concern just for researchers but for all those who engage in public spaces of political discourse.

Participant's interpretations of speech indicated that they viewed speech as fundamental to their relationships both online and offline. This interpretation, as discussed in Section 3.5.5, appears to differ from the culturally pervasive US-interpretation of speech. This raises the question: do platforms share the same perceptions of free speech and speech as their users? This question motivates the work presented in Chapter 4.

In the following chapter, I draw on the findings of this work to closely examine the presence of the US-interpretation of speech and free speech in the speech moderation practices of present-day social media platforms. Specifically, I do a close reading of X's Violent Speech moderation policy and Meta's Community Guidelines on Violent Speech.

Table 3.1: Participants in Interview & Diary Studies

P#	Age	State (abbr)	Race / Ethnicity	Location Setting	Highest Level of Education	Social Media Platforms	Active Social Media Use	Interview or Diary Study?
P1	25	MA	African American	Urban	B.S.	Instagram, Reddit, X, YouTube	12 years	Interview
P2	21	PA	Asian	Urban	High School	Instagram, TikTok, SideChats	9 years	Both
P3	25	CA	South Asian	Urban	B.S.	Instagram, Facebook, SnapChat	10 years	Both
P4	26	ME	White Ashkenazi Jew	Rural	M.A.	Instagram, Reddit, Facebook	15 years	Both
P5	26	MO	White	Urban	B.S.	Instagram	12 years	Interview
P6	25	NJ	Asian	Suburban	M.A.	Instagram	7 years	Both
P7	21	GA	N/A	N/A	N/A	Instagram, TikTok	N/A	Interview
P8	24	NY	South Asian	Urban	B.S.	Instagram, TikTok, SnapChat	11 years	Interview
P9	24	MA	Southeast Asian	Urban	B.S.	Instagram, X	9 years	Interview
P10	25	NJ	South Asian	Suburban	B.S.	Instagram, TikTok, SnapChat	12 years	Interview
P11	25	MA	Asian	Suburban	B.S.	Instagram, X	9 years	Both
P12	25	GA	N/A	N/A	N/A.	Instagram	N/A	Interview
P13	25	NJ	Asian American Korean	Suburban	M.A.	Instagram	13 years	Both
P14	25	NJ	Asian Indian	Suburban	M.A.	Instagram, TikTok	10 years	Interview
P15	19	GA	N/A	N/A	N/A.	Instagram	N/A	Interview
P16	25	VT	Asian American	Suburban	B.S.	Instagram	12 years	Both
P17	18	PA	Asian Indian	Urban	High School	Instagram, TikTok	9 years	Interview

Table 3.2: Participants in Interview & Diary Studies - Continued

P#	Age	State (abbr)	Race / Ethnicity	Location Setting	Highest Level of Education	Social Media Platforms	Active Social Media Use	Interview or Diary Study?
P18	23	GA	Asian Indian	Urban	B.S.	Twitter/X, Instagram, WhatsApp, YouTube	6 years	Diary Study
P19	27	GA	White Middle Eastern	Urban	M.A.	Instagram	11 years	Diary Study
P20	26	MA	Indian	Urban	B.S.	Instagram	14 years	Diary Study
P21	25	GA	Asian Japanese Malay-Chinese	Suburban	B.S.	Instagram	12 years	Diary Study
P22	25	D.C.	South Asian	Urban	B.S.	Instagram	10 years	Diary Study
P23	22	NH	White	N/A	B.S.	Instagram	9 years	Diary Study

CHAPTER 4

SOCIAL MEDIA SPEECH MODERATION POLICIES & FREE SPEECH

4.1 Introduction

This chapter broadly examines the underlying concepts of 'free speech' that shape speech moderation practices on social media platforms in the U.S. Specifically, I ask, what interpretations of 'free speech' underlie social media speech moderation policies, and how do they shape moderation practices?

In this chapter, I present two close readings of two violent speech moderation policies to examine their underlying interpretations of 'free speech'. I choose policies focused on mitigating violent and threatening speech because, in the United States, such speech is not considered 'free' and not protected by the First Amendment [28]. As discussed in Chapter 2, although social media corporations do not need to abide by the US-interpretation of free speech, many choose to do so. Since violent and threatening speech is one of the few forms of speech not protected by the First Amendment, social media platforms may approach the moderation of such speech with more rigor and specificity compared to speech that is not protected by the First Amendment, such as hate speech.

Through the two close readings, I illustrate how the pervasive US-interpretations of 'free speech' shape the design of a social media speech moderation policy, specifically in interpretations of 'violent speech' and 'threatening behavior', intentionality, and credibility. I argue that the US-interpretations of speech underlying these policies may broadly influence the limited moderation practices of hateful and violent speech on these platforms as this notion of speech views speech as 'just words' and does not account for the impact and context in which the speech was made.

Throughout this chapter, I will be referring to an individual who makes a text-based

social media post as a 'speaker', and an individual who is targeted by the post as a 'target', 'recipient', or 'intended audience'. I use these terms to draw a direct connection between the literature on free speech and moderation policies.

The first close reading is of X's Violent Content policy, and the second is of Meta's Violence and Incitement Policy. I focus on these two platforms due to the large amounts of political discourse that occurs on X, Facebook, and Instagram. These platforms have enabled significant public political discourse in multiple global elections [51, 212, 213] and between politicians and their constituents [67].

Both policies were analyzed in October 2024. X's Violent Content policy, at the time, was last updated in May 2024. Meta's Violence and Incitement policy, at the time, was last updated in September 2024. These policies may have been revisited in anticipation of the 2024 U.S. Presidential election, which took place in November 2024. Both platforms have been spaces of political advertising, discourse, and news in previous elections. Both platforms have also been widely criticized for enabling election mis/disinformation, and hateful and violent speech targeting candidates in past elections [212, 27, 25]. Furthermore, October 2024 marked one year of the Israel-Gaza conflict, which has been at the forefront, for many, of political discourse in the US [214].

4.2 Close Reading 1 - X's Violent Content Policy

This close reading is of X's Violent Content policy as of October 2024 [70]. The purpose of this policy is to outline the types of content that X identifies as 'violent' and the actions the platform will take against such content.

4.2.1 Overview of Policy

X's Violent Content Policy contains two primary sections. The first section provides an extensive explanation for how 'Violent Content' is defined, what types of violent content are allowed, what types of violent content will have restricted reach, and what types of

violent content are not allowed. X's Violent Content Policy separates violent content into two categories: Violent Speech and Violent Media. Violent Speech refers to text-based content that contains references to violence. Violent Media refers to visual-based media (image, video) depicting violence.

The second section of the policy includes guidance on how to label content appropriately for users who would like to post graphic content on X. Users posting graphic content must adjust their settings to add content warnings to their posts. This policy was last updated in February 2025, however, the analysis of this policy was conducted in October 2024.

For this close reading, I focus on the Violent Speech-specific policy, as my work broadly focuses on interpretations of 'speech'. In the following sections, I outline the content of the policy and analyze the policy's interpretation of 'violent speech'.

4.2.2 Moderating Violent Speech on X

Interpretation of Violent Speech

X's section on violent speech begins with the statement, 'We prohibit Violent Speech that we consider high in severity and the likelihood of harm', indicating that the moderation system has a severity threshold that is consulted when making moderation decisions. However, no further explanation is provided regarding this threshold.

The policy identifies four primary examples of violent speech, 1) violent threats, 2) wishes of harm, 3) incitement of violence, and 4) glorification of violence. According to the stated policy, 'violent threats' are statements that 'threaten to inflict physical harm' on a person, homes, and or public infrastructures such as bridges or buildings. 'Wishes of harm' are statements in which the speaker expresses a desire or wish for their recipient to experience some form of physical harm, whether through illness or violence. Statements that begin with 'I hope' or 'I wish' may be included in this. 'Incitements of violence' refer to statements in which the speaker provokes or encourages others to commit acts of

physical harm. 'Glorification of violence' refers to statements in which the speaker may praise or applaud acts of physical violence or harm against individuals or animals.

The policy states that speech that is violent but is made in a situation with 'no clear abusive or violent context' will be allowed on the platform. Specifically the policy states,

> We seek to evaluate and understand contextual nuances behind the conversation before taking action, and we allow expressions of Violent Speech when there is no clear abusive or violent context, such as:
>
> - hyperbolic and consensual speech between friends, or during discussion of video games and sporting events;
> - figures of speech, satire, or artistic expression when the context is expressing a viewpoint rather than instigating actionable violence or harm;
> - quotes from books and movies, music lyrics or poetry when the context is expressing a viewpoint rather than instigating actionable violence or harm.

For instance, if the post is considered an instance of hyperbolic speech or satire, it may remain on the platform. The discussion of 'context' indicates that the moderation system seeks to understand the factors of the situation in which the speech was made. It remains unclear how the moderation system seeks to understand what constitutes a 'clear abusive or violent context'.

X's Violent Content policy has a narrow interpretation of 'violent speech'. The policy interprets violence based on the inclusion of any reference to physical harm. The policy does not further clarify whether instances of other forms of harm such as financial, emotional, and or mental harm are actionable under this policy. X's policy does not provide specific phrases or terms that may aid in identifying what statements indicate violence, and does not elaborate on what criteria will be used to assess a statement as 'violent'.

X's policy roots its interpretation of 'violent speech' in a reference to 'physical harm'. This indicates that X's interpretation of 'violent speech' is dependent on whether the statement references the specific actions the speaker may take. For instance, the statement, 'I will ruin your life' would not be considered 'threatening' as stated by X's policy, as there is no reference to physical harm. In this statement, the speaker has not explained how they may take action to ruin another individual's life, and whether their course of action includes physical harm. Thus, this statement does not meet the criteria of X's policy.

Policy Actions

The platform states that should content violate this policy, it will be removed. In some cases, depending on the severity of the violation, the content will be made less visible on X or the account will be suspended. X users may appeal if they disagree with the action taken against them.

Statements in immediate violation of the policy will be removed. This includes statements that fall into the four main categories of Violent Speech: violent threats, wishes of harm, incitement of violence, and glorification of violence. Those who violate this policy multiple times may see their posts having limited reach or will have their account suspended.

The policy describes instances in which certain content containing violent speech may remain on the platform but have limited visibility across the platform. For instance, this may occur when the harm the speech may cause is minimal or 'non-deliberate', if the speech was made in self-defense, if there is no clear intended recipient, if it was targeting a 'perpetrator of major harm', or if it used 'coded language'. There is no further explanation for how the moderation system determines these specifics. Additionally, it is unclear which limited set of X users may come across this content or how the 'limited visibility' is determined.

X's Violent Content policy does not elaborate on the actions that may be taken if a

user repeatedly violates the policy. The policy does indicate that users may be suspended for posting content that contains violent speech or media, but the length of suspension is unclear. The policy does not indicate that users will be banned entirely from the platform for violating the Violent Content policy.

4.3 Close Reading 2 - Meta's Violence and Incitement Policy

This close reading examines Meta's Violence and Incitement policy that is implemented on both Facebook and Instagram. This close reading was conducted in October 2024, prior to the changes made to Meta's moderation system in January 2025 [32]. This policy outlines how Meta interprets 'violence' and 'incitement of violence' and how the company addresses such content on both Facebook and Instagram.

4.3.1 Overview of Policy

Meta's Violence and Incitement Policy, last updated September 27th, 2024, consists of two sections. The first section lists the types of content that the platform removes, and the second section lists content that the moderation may need more context about before removing. The policy also includes links to previous iterations of the policy to highlight how the policy has changed over time. Sections that have been updated or edited are highlighted in previous iterations of the policy. For instance, between the August 13th version of the policy and the August 29th version of the policy, the following additions were made to the list of types of content that require more context for moderators to take action:

> Threats of high- or mid-severity violence in the defense of self or another human when the criteria below are met.
>
> - Against a person (excluding persons identifiable by name of face, people targeted based on their protected characteristics, and children)

- In the context of home entry or interpersonal violence that is proportional to the violence responded to and is an immediate threat
- The potential impact on voice outweighs the risk of imminent violence

4.3.2 Moderating Violent Speech on Meta's Platforms

Interpretation of Violent Speech

Meta's policy [71] begins by providing the reader with an explanation of why such a policy exists in the first place. The policy draws parallels between online speech and offline behavior, indicating a broader understanding of how instances of unmitigated online violent speech and behavior on platforms such as Facebook and Instagram can severely impact the lives of people offline. As such, this policy exists to prevent instances of offline violence. Despite the policy being labeled as 'Violence and Incitement', much of the policy focuses on 'Threats', with some references to 'Other Violence' which includes statements that call for someone to commit a crime, references to domestic violence, and or 'instructions on how to make or use weapons…'. This suggests that the policy interprets most violent speech as 'threatening behavior', but does not consider references to 'domestic violence' or 'building a weapon' as 'threatening behavior'.

The policy states

> While we understand that people commonly express disdain or disagreement by threatening or calling for violence in non-serious and casual ways, we remove language that incites or facilitates violence and credible threats to public or personal safety.

The policy further explains,

> In determining whether a threat is credible, we may also consider additional information such as a person's public visibility and the risks to their physical safety.

Meta's Violence and Incitement policy explicitly states that threats made in a 'condemning' context, in which they are targeting terrorist groups, or in an 'awareness-raising' context, are not actionable under this policy. The policy further states that threats made in a 'sports' context, such as during an intense sports game, are not actionable under this policy. Meta delineates between the types of content the policy immediately removes and the types of content that require more information and context before making a decision. These decisions are based on how 'severe' the content is deemed to be.

Content with a 'high to mid-severity of violence' is immediately removed from the platform. 'High-severity violence' is characterized as violence that may result in death. 'Mid-severity violence' is characterized as violence that may result in serious injury. Statements with mid-severity violence are only removed if they are deemed 'credible'.

Meta's policy also explicitly states that separate action is taken on statements of violence targeting individuals made on the basis of 'protected characteristics' which are: "race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease" [71]. The policy states,

> In addition to the universal protections for everyone, all private adults (when self-reported), children and persons or groups of people targeted on the basis of their protected characteristic(s), are protected from threats of low-severity violence.

These individuals are also protected against statements with high and mid-severity violence. Furthermore, statements of violence targeting individuals based on protected characteristics, children, or targeting law enforcement and election officials, are removed regardless of credibility.

Forms of violent content that require additional information or context before action is taken are 'coded statements', threats with high or mid-severity violence that are made in self-defense or defense of someone else, statements implying the potential for violence in public spaces, and specifically,

> Claims or speculation about election-related corruption, irregularities, or bias when combined with a signal that content is threatening violence.

'Coded Statements', as defined by Meta are "where the method of violence is not clearly articulated, but the threat is veiled or implicit". In such statements, the lack of information and clear articulation is purposeful as to employ plausible deniability [215, 216]. According to the policy, a coded statement is removed if it contains both a 'threat signal' and a 'contextual signal'.

There are four 'threat signals' identified by the policy. One, the statement references historical or fictional acts of violence. Two, the statement suggests that the speaker knows sensitive and or personal information regarding the recipient that could result in harm if it is exposed. Three, the statement is 'shared in a retaliatory context'. For instance, statements made where the speaker feels they have been wronged and want to retaliate in a violent manner. The final threat signal is a statement that serves as a dog whistle or an indirect call to action to commit violence.

There are three 'contextual signals' identified by the policy. One, the statement is targeting a child. Two, the targeted recipient or an authorized representative reports the statement, or three, a local context expert determines that the statement could be interpreted as violent and or could incite violence. If a statement contains both a threat signal and a contextual signal, it will be removed from the platform.

Similar to X, Meta's interpretation of 'threatening behavior' focuses on threats of violence, and more broadly 'physical violence'. As aforementioned, the policy characterizes references to domestic violence and 'building weapons' as 'Other Violence' and not necessarily threatening behavior, suggesting that unless a statement contains explicit indications of an intention to commit or wish for physical harm, the statement is not necessarily 'threatening'. Furthermore, the hierarchical ranking of violence (high-severity, mid-severity) assigns a value to violent content. High-severity violent content is defined by its inclusion of explicit references to physical violence and is the type of content that is immediately

moderated.

Policy Actions

Meta's policy indicates that content removal is the main action Meta takes when content is identified as 'violent'. The policy also explains that the moderation system will seek additional context and information if necessary before making a decision. This occurs through a combination of an AI-driven system and a team of human reviewers. The platform provided more insight into this process in a 2022 blog post, explaining, "...our technology [the AI-driven system] might be unsure whether a post contains bullying, a policy area that requires extra context and nuance because it often reflects the nature of personal relationships. In this case, we'll send the post to review teams that have the right subject matter and language expertise for further review" [217]. From this explanation, it appears that Meta's policy seeks to understand the context in which the speech was made by gathering information on the relationship between both parties. The Violence and Incitement policy additionally explains that a local context expert is sometimes consulted to confirm or deny whether a statement could be interpreted as violent or inciting violence. However, one can only speculate how this occurs in practice.

4.4 Summary of Close Readings

Both X's and Meta's policies on the moderation of violent speech suggest a limited interpretation of speech that is considered violent and or threatening. These policies emphasize 'physical' forms of harm, insinuating that violent statements will be moderated if they contain explicit references to physical actions made by the speaker that could harm another individual. This demonstrates a speaker-centric approach to moderation, in which the policies identify speech to moderate based on how direct the speaker is in their intention to commit or wish for harm. Statements that are more explicit in their intentions to wish for or to commit harm are considered more credible and therefore 'worthy of moderation'. The

speaker-centric approach disregards a recipient's interpretation of violent speech targeting them. These policies do not consult or take into account a recipient's perspective when determining how to moderate an instance of violent speech, suggesting that their perspective is not as important or relevant to the 'credibility' of the violent nature of the speech. This speaker-centric approach to moderation is individualistic and contains underlying Libertarian concepts of 'speech' and 'free speech'. I elaborate on this further in the Discussion.

4.5 Analysis of the Policies

Both X and Meta center the speaker as part of their policies on violent speech, illustrated through the policies' emphases on 'intention' and 'credibility'. Specifically, both policies indicate that the inclusion of an explicit intention to commit or wish for harm is a prerequisite to taking action against the statement. Intentionality and credibility are correlated, as having an explicit speaker intention to commit or want to commit harm within a statement, makes the statement appear more credibly violent or threatening. A statement that is considered more 'credibly threatening' will presumably be more likely to be moderated.

X and Meta's policies focus on the speaker over the recipient or target of the statement, demonstrating a moderation approach that favors a speaker's ability to express a statement of violence over the recipient's feelings of safety. This is a Western notion of 'free speech', one that is rooted in ideas of individualism and disregards the effects or consequences of speech on others. In the following sections, I explain X and Meta's focus on intention and credibility, drawing connections between these concepts and the US-interpretation of free speech. I then discuss the potential implications of engaging this US-interpretation of free speech within a social media platform context.

4.5.1 Intentionality

X's Violent Content policy and Meta's Violence and Incitement policy are designed to address content that contains explicit indications that the speaker intended to commit harm.

In both policies, this is shown through an emphasis on addressing speech that includes references to 'physical harm', either through an explicit wish for direct physical harm to occur or an explicit intention that the speaker will commit some form of physical harm. Although X and Meta do not provide examples of 'intention', linguistic scholarship on threatening statements finds that intention is often expressed through phrases such as 'I will' or 'I am going to' [218, 219].

The policies' focus on speaker intention illustrates a framing of speech that gives preference to the speaker over the recipient of the speech. Language philosopher Jennifer Hornsby argues that this framing of speech is Libertarian, and in this view, a speaker is not at fault for any "harm" they cause the recipient. From this perspective, speech is offensive only if the recipient believes that the speaker intended for them to believe that the speech was offensive [211]. In other words, if the recipient is offended or "harmed" by words that the speaker said, it is the recipient's fault that they understood it that way. This, in turn, may result in implicit or indirect forms of violence and or threatening behavior targeting individuals to be disregarded or not considered 'threatening', forcing these individuals to either self-censor or be excluded from discourse out of fear of what may occur.

X's Policy

X's Violent Content policy does not explicitly address 'intent' or 'speaker intentionality' when defining violent speech. However, as aforementioned, the policy suggests that only statements that directly express a wish for physical harm, threaten to commit physical harm, encourage others to commit physical harm, or praise others who have committed physical harm will be actioned upon.

The policy's emphasis on statements referencing physical harm serves as a requirement for direct references to intentionality. If a statement references physical harm, it references the actions that the speaker may take themselves, or want someone else to take, thus indicating an intention to harm the individual being targeted. For instance, consider the

statement, “If only you were hit by a bus…”. This statement contains a direct reference to an act of physical harm (“hit by a bus”). The speaker does not implicate themselves as the one to want to commit the harm, because they do not explicitly say that they want to be the ones to hit the targeted individual with the bus. However, they do implicate themselves as wanting the recipient of the statement to be hit by the bus. Thus, such a statement would potentially be considered a ‘wish for harm’ which is actionable under X’s Violent Speech policy.

X’s policy further emphasizes its requirement for direct references to intentionality in its explanation that statements with ‘coded language’ and statements with unclear targets can remain on the platform but will have limited visibility. For example, a statement with a ‘dog whistle’ may be interpreted as threatening by a recipient, but will not be removed regardless of how a recipient may feel being targeted by the statement. Linguists call such statements, ‘indirect threats’ as they do not contain direct references to intentionality [216]. Linguist Roger Shuy gives the example of the statement, ‘How’s the leg?’ as an indirect threat. Such a statement appears harmless and does not implicate the speaker as intending to commit harm. But suppose the speaker of the statement has already physically harmed the recipient. Given this prior experience between both parties, this statement may be interpreted as sinister and potentially threatening [215]. However, the words included in the statement do not illuminate the underlying context and relationship between both parties. Thus under X’s policy such a statement would not be removed but may have limited visibility. The lack of explicit speaker intention within such a statement, despite how threatening the statement may be considered, diminishes how credible it is considered. I elaborate on the concept of ‘credibility’ further in the following section (Section 4.5.2).

X’s policy on Violent Speech, and Violent Content more broadly, suggests that online speech that does not explicitly indicate that the speaker intended to commit an act of physical harm or wish for physical harm to occur on another individual, is not an instance of violent speech. While a statement may be actioned upon by other platform policies such as

'hateful conduct' [220] or 'abuse and harassment' [221], the speaker's intention is centered in X's policies to determine how to mitigate the statement. In doing so, the policy minimizes the role that the recipient's interpretation or perspective of statements targeting them plays in the process of determining whether the statement should be moderated, and even more broadly, considered 'violent'.

Meta's Policy

Similar to X's policy on Violent Content, Meta's policy on Violence and Incitement does not explicitly reference 'intent' or 'speaker intention', but engages a narrow interpretation of 'violent speech' to limit the speech that is moderated under this policy. Meta's policy highlights 'threats of violence' and focuses on identifying content that 'poses a genuine risk of physical harm or direct threats to public safety'. The emphasis on 'physical harm' and the limited interpretation of 'violent speech' suggests that the policy, similar to X's policy, centers on the speaker. Furthermore, the process by which moderators understand what is considered 'a genuine risk' is unclear. One may speculate speech with explicit speaker intention may be considered 'a genuine risk', whereas speech without explicit speaker intention may not be.

Meta's policy also states that 'coded statements' where the 'method of violence is unclear' may be removed if these statements have both a 'threat signal' and a 'context signal'. There is no indication that the person being targeted is consulted to potentially clarify the coded statement. I elaborate on this in the section below, however, the inclusion of this information in the policy indicates that the context in which the statement was made is considered, but only partially.

Unlike X's policy, Meta's policy does make multiple references to understanding the context surrounding the speech before making a moderation decision. The policy explicitly draws attention to certain populations who may be considered under alternative rules such as children or an individual who has been targeted based on characteristics such as their

race, gender identity, or disability. The policy also claims to examine 'local context' in instances where it may be necessary, to understand whether a statement could be considered violent or not. However, the target or recipient's perspective is not considered as part of this context. From these claims made in the policy, Meta's policy may be centering the speaker, but not necessarily ignoring the speaker's connection or relationship to the individual being targeted by the speech.

In both policies 'intention' and 'speaker intention' is not explicitly mentioned, but indirectly referenced. Both policies are oriented towards speakers/creators, explaining to them what is allowed to be said on the platform and what is not. Both policies emphasize moderating content with direct references to physical harm and that contain 'genuine risk', such as speech that directly implicates the speaker as the one to commit, or want to commit an act of physical harm.

4.5.2 Credibility

References to explicit speaker intentionality to commit harm lend to advancing a statement's 'credibility', and the concept of 'credibility' is consistently referenced in the context of speech moderation. As aforementioned in Chapter 2, social media platforms do not need to abide by free speech laws, however, the law continues to shape their approach to the moderation of speech. Although both X and Meta do not clarify what they mean by 'credible', the term 'Credible threat' has often been used in the law context. Sociologist Sarah Sobieraj defines a threat as 'credible' "... if the target might reasonably believe the threat is serious, could feasibly be carried out, and that the danger is imminent" [67]. She further explains that the term 'credible' is used as a threshold for the law to hold someone accountable for making a threat. The 2023 U.S. Supreme Court (SCOTUS) ruling in the case *Counterman v. Colorado* further elaborates on what constitutes a 'credible threat' in an online context in the United States [222]. The Court's ruling explained that should an individual believe that their statement might be interpreted as threatening, and make the

statement anyway, then the statement would be considered 'credibly' threatening and not protected as free speech. But should an individual not believe their statement could be interpreted as threatening, that statement would be considered free speech. One's intention behind their statement shapes how 'credible' the statement is considered.

While Sobieraj's definition of 'credible' draws attention to the target's perspective as fundamental to defining a statement as 'credibly threatening' or not, the U.S. Supreme Court ruling does not align with this definition. The SCOTUS ruling centers the speaker and their 'stated' intention to determine how credible one should interpret the 'threat' to be, rather than the victim and their interpretation of the statement [223, 224].

This method of identifying what is 'credibly threatening' further reinforces the underlying individualistic notions of speech in the United States. This method, as explained by the 2023 SCOTUS ruling, enables speakers to plausibly deny any knowledge that their statement might be interpreted as threatening, and instead push the blame of interpretation onto the target recipient. As mentioned previously, language philosopher Jennifer Hornsby argues that through this interpretation of speech, the speaker can blame the hearer for interpreting their speech as 'threatening' or 'harmful' [211].

As discussed in the previous section, X and Meta's policies similarly follow the model of interpreting threatening behavior laid out by the 2023 SCOTUS ruling, despite not needing to abide by U.S. free speech laws and interpretations.

X's Policy

X's policy explicitly mentions that statements with 'coded language' or those that do not have explicit intentionality can remain on the platform but will have limited visibility. Since such statements obfuscate their expressions of violence or do not explicitly target their recipient, they may not be deemed credible and are therefore not removed. Additionally, statements with references to broad instances of violence that do not specifically target other individuals will not be removed but may have limited visibility. The lack of specificity

regarding who is being targeted may minimize how credibly violent the statement may be. For instance, the statement 'We have guns.' may not be considered an instance of violent speech under X's policy, as there is no reference to physical harm, the target of the statement is unclear, and the statement does not implicate the speaker as wanting to commit any harm, despite the reference to a physical weapon.

X's Violent Content policy states that hyperbolic speech, quotes from media, figures of speech, satire, or artistic expression that contain references to acts of harm or violence is not considered 'violent speech', as these references to violence may not be considered likely or plausible. This suggests that the plausibility of the violence referenced in the statement influences its credibility as 'violent speech' and whether it will be moderated as such.

The specific inclusion that 'artistic expression' and quotes from media in the description of content that may not be actionable under this policy, puts the policy at odds with the 2015 SCOTUS opinion in *Elonis v. United States*. In *Elonis v. United States*, the defendant, following a separation from his spouse, posted rap lyrics on Facebook depicting violence against a former lover. While the defendant argued that he was simply posting rap lyrics and he had no intention of harming his spouse offline, his spouse felt threatened and took legal action. In this instance, the ambiguity of the rap lyrics may have disguised the intention of the defendant, and have enabled them to hide their intention under 'artistic expression' [225]. While the Supreme Court ruled in favor of Elonis, due to an error regarding how the case was presented to the jury, they did make a judgment regarding how online threats should be considered. The majority opinion stated that for a statement to be considered a threat, the speaker needs to have the intent of making the threat and know that what they post will be interpreted as a threat. As historian Angus Johnston succinctly summarized in his analysis of the Supreme Court opinion, "If you make a threat online, and you know the person who receives it will see it as a threat, you're guilty of violating federal law. It doesn't matter if you claim that it's protected speech, or put a smiley face at the end, or

point out later that the threat was really just lyrics from an old Beatles song. If you send a threat and you know it'll be interpreted as a threat, you're guilty. Period" [226].

The SCOTUS opinion further explained that online posts are typically made for a wide audience, so if it is targeted, it will be taken more seriously [225]. The very nature of X, and social media in general, encourages sharing one's thoughts with a public audience, not necessarily directly addressing one individual. In this context, users can technically make threats on social media without their targets knowing, by simply not tagging them or not mentioning them. Thus, when a user does make a statement about harming an individual and then mentions that person's account, they are ensuring that the tweet will be seen by the person mentioned. A person will take a threat more seriously if they were specifically intended to see it [227].

Deeming instances of violent artistic expression as not 'credibly threatening' may enable speakers to make targeted violent speech that is disguised and or only known to the speaker and the intended recipient.

Meta's Policy

Meta's policy repeatedly mentions credibility as part of their moderation approach. The policy states that the platform will remove statements that are 'credible threats', and to determine credibility, moderators will seek out additional information such as, "a person's public visibility and the risks to their physical safety." Furthermore, while the policy states that it will consider statements that are 'coded' or 'imply violence' if these statements contain certain 'threat' and 'context' signals, the policy does not indicate that moderators seek out the recipients' perspective on whether they felt the statement targeting them was threatening. Meta's policy does explicitly state that 'credibility' will not be accounted for when an expression of physical harm is targeting children or an individual based on their protected characteristics.

The policy's defined 'levels of violence' further support the platform's emphasis on

credibility. Statements with 'high-severity violence' are immediately removed, whereas statements with 'mid-severity violence' are removed if deemed 'credible'. Meta suggests that credibility is determined through identifying context and information regarding the target's safety. However, Meta does not explain how moderators determine what is considered 'high-severity', nor how it differs from statements of 'mid' or 'low' severity.

The hierarchical ranking of violence is a categorization mechanism developed by forensic linguists to determine what types of threats should be immediately acted upon [228, 229, 230]. Forensic linguists have created hierarchies categorizing what forms of threats would be considered "high-risk" and what would be considered "low-risk". Statements that are more outlandish and potentially hyperbolic in nature are considered low-risk, whereas statements that are more specific in detail and are seemingly more probable are considered high-risk [229]. For example, the statement "I will push you off a cliff" would be considered low-risk due to the relatively improbable nature of the form of violence referenced. In contrast, the statement "I will bring a gun to your house and shoot you" would be considered high-risk due to the degree of specificity of details mentioned and the form of violence is potentially more probable in the United States.

4.6 Discussion

4.6.1 Slippery Slope & Chilling Effects

As discussed in Chapter 2, the US-interpretation of free speech draws on libertarianism, which emphasizes individualism and individual rights. As such, free speech is protected because it is fundamental to protecting one's autonomy and individual liberty [81, 76, 231]. This interpretation of free speech protects instances of hateful and violent rhetoric under the guise of protecting one's right to individual liberty. Legal scholar Mary Anne Franks argues that the Civil libertarian perspective of free speech justifies the protection of hateful and violent speech due to concerns of a chilling effect or silencing effect on speakers [66].

In the context of U.S. law, a chilling effect refers to the concept of people choosing not

to speak out for fear of facing governmental or legal repercussions. Civil Libertarians argue that should a policy be put in place to mitigate hateful speech it would cause a chilling effect [64, 232]. Franks is critical of the Civil Libertarian interpretation, arguing that the concern of a chilling effect enables Civil Libertarians to distance themselves from the hateful and violent speech that is being defended and 'to characterize this defense as the only thing keeping good speech from being silenced tomorrow' [66, 65]. In other words, the concern of chilling effects and silencing on speakers is often used to justify the protection and defense of hateful and violent speech because, if one starts to moderate or censor hate speech it could set a precedent for the moderation and censorship of other forms of speech as well. Franks refers to this as the 'slippery slope' argument.

Platforms such as X and Facebook have previously discussed the potential of a chilling effect occurring through moderation practices to address hateful and violent speech online. These platforms have expressed concern about users 'weaponizing' a hate speech or violent speech policy to wrongfully accuse someone of making hate speech or violent speech, and therefore 'chill' speech [3, 233, 234]. These platforms are concerned regarding the chilling or silencing effects on speakers, in which a speaker is worried about facing repercussions such as suspension, or content removal, for expressing their thoughts. However, they are disregarding the potential chilling or silencing effect on recipients or witnesses of speech that is violent and or hateful. Those who choose to self-exclude or disengage from spaces due to their fear of expressing their opinion and being targeted by violent speech and or threatening behavior in response. Those such as the young adult Gen Z women who participated in the research work discussed in Chapter 3. The young adult Gen Z women participants I encountered cited their fear of facing negative social, physical, and professional consequences for expressing their opinions in political spaces as a primary reason for disengaging from those spaces.

X and Meta's policies suggest that they engage the civil Libertarian approach to speech and similarly disregard an intended recipient's perspective of the speech when making a

moderation decision. The policies are in place to aid the speaker in potential acts of silencing, but not the recipients who may be silenced. X and Meta's lack of consideration for 'coded statements' or statements that imply violence, their hierarchical perspective on the severity of violence, and their evaluation of how 'credibly violent' the statement is suggests a moderation approach that makes it difficult to moderate all statements perceived as potentially violent and or threatening by intended recipients. Furthermore, these policies provide avenues for speakers to appeal moderation decisions but provide limited recourse to recipients of hateful and or violent speech. Feminist activist Bailey Poland writes about how many women do not have the resources to adequately address threatening behavior targeting them online, often leading them to ignore the behavior [61].

X's Violent Content policy states that a threat of violence will be removed, and at most, the speaker will be suspended. Following this suspension, the speaker may once again target their intended recipient with threats of violence. The policy does not indicate that repeat violators will be banned from the platform. Nor does the policy indicate any protections for intended recipients against repeat violators.

From the stated policy, if an individual is continuously being targeted with violent speech, they should block the account, continuously report the account, engage in a back-and-forth with the account, or leave the platform. These actions do not provide much protection for the recipient, particularly should the speaker shift from online threatening behavior to offline. Additionally, there are limited forms of recourse provided to people offline regarding threatening online behavior. In one example that Poland provides, she discusses the experience of a woman attempting to file a police report regarding threats of violence targeting her on Twitter. The woman felt that her local law enforcement did not understand the nature of repeated threatening behavior online because, from their perspective, threats are ephemeral, especially if they are online. Thus, they cannot take much action [61].

This interpretation of threatening behavior appears to exist both online in the policies

discussed, and offline in the practices of law and law enforcement. This interpretation views speech as 'just words' that carry no further implications on the lives of those who receive it or are witnesses to it. It disregards the mental, physical, and emotional toll that being targeted by violent and or threatening behavior online can have on a person.

4.6.2 Speech as Just Words

In centering the speaker and focusing on intention and credibility, platforms may be subscribing to the Libertarian interpretation of speech that speech is 'just words'. U.S. interpretations of free speech disregard that speech has an impact beyond its utterance, and is therefore 'just words'. Engaging this individualistic interpretation of speech may enable platforms to distance themselves from the consequences and harm caused by these words. For instance, in 2018 political analyst Rochelle Ritchie publicly discussed Twitter dismissing her report of threatening tweets she received. One tweet she received stated, 'We will see u 4 sure. Hug your loved ones real close every time you leave home.' and included a photo of her. At the time, Twitter dismissed her report, stating that the tweets did not violate their policies. Two weeks after dismissing her report, Twitter took note of the severity of the tweets when the individual making the tweets was arrested for mailing pipe bombs to various prominent U.S. political leaders [235, 227].

Unfortunately, such an example is one of many in which individuals vocalized their concerns regarding their safety online but were told that the content or speech targeting them did not violate any platform policy. Their voices were effectively unheard. Although platforms have a difficult responsibility to mitigate hateful and violent content, their individualistic approach to moderation is partial to the speaker and ignores the recipient of the speech.

Furthermore, the view that speech is 'just words' enables platforms to shift the burden of protecting oneself from online harm to the individual who was targeted. As philosopher Jennifer Hornsby notes, under this notion of speech, an individual may be blamed for their

'misinterpretation' of speech targeting them [211]. As aforementioned in the previous section, if speech is considered 'just words' there is no need to provide avenues of recourse for recipients of violent and hateful speech because the speech does not cause harm or offense. Individuals who have been targeted are left to identify mechanisms by which they can adequately deal with the speech targeting them. For example, users block and report toxic behaviors targeting them, and in some instances maintain records on instances of cyberstalking and cyberharassment to report to law enforcement [61]. The burden of maintaining a space for them to engage in discourse is pushed to those who are often targeted and marginalized from these spaces online. In some cases, as evidenced by findings in Chapter 3, many users choose to disengage from the space entirely as a mechanism to protect oneself from online harm.

Additionally, engaging the speech as 'just words' interpretation may enable platforms to rely more substantially on automated systems. In recent years, both X and Meta have utilized automated systems to identify toxic speech [32, 236]. The speech as 'just words' interpretation, which views speech as not having an impact beyond the specific utterance, may underlie these automated systems that similarly view speech as discrete words. An automated system may be trained to identify specific phrases of violent intent or physical harm and address speech containing those phrases accordingly. The Libertarian view of speech as 'just words' may serve to enable more reliance and engagement with automated systems that do not account for the broader context in which the speech was made and to whom it was made. This may, in turn, exacerbate the uneven application of moderation systems and policies [237].

Hate speech and violent speech affect people differently, depending on their context and background. For instance, terms such as 'slut' and 'bitch' have a long history of being incredibly offensive to women, but in recent years have been reclaimed as a word that can be used between friends [63, 238, 239]. The context of how the term is used and who is using it affects the interpretation of the term. However, in engaging the speech as

'just words' interpretation, words such as 'slut' and 'bitch' may be considered offensive regardless of context. Under such an interpretation, the moderation of specific words may be reduced to a binary 'yes or no' decision.

4.6.3 Interpretations of Free Speech and Gen Z Women's Engagement

The findings presented in Chapter 3 illustrate how the young adult Gen Z women participants have chosen to disengage from political spaces online. They are critical of social media being used as a space to discuss politics, are skeptical of the political information they see online, and express a fear of facing social, physical, and professional repercussions for expressing their political opinions online. While the participants of this work do not explicitly state that content moderation practices affect their forms of engagement in public discourse, they do attribute online toxicity as a reason for disengaging.

Content moderation practices serve as the primary mode of mitigating online toxicity among platforms, but as discussed throughout this chapter, their underlying US-interpretation of free speech prevents the thorough mitigation of toxic speech and behavior and instead continues to platform a large portion of it. In the work presented in Chapter 3, participants spoke of seeing racist, misogynistic, and xenophobic content, which deterred them from expressing themselves in these spaces. Much of this content is enabled to be on X, Facebook, and Instagram as it is offensive yet still considered legitimate discourse. This content may be flagged by a 'hateful conduct' or 'abusive behavior' policy, but those who make these posts are enabled to continue doing so. These platforms take a lax approach to mitigating content that does not overtly threaten, particularly in the context of political discourse.

Observing users go unpunished for posting toxic content may negatively affect one's perception of these spaces. Users targeted by this racist, misogynistic, and or xenophobic content may feel that these spaces are not for them. With regard to women in these spaces, prior research has shown how the lack of adequate mitigation of toxic behavior online has

broadly affected women's perceptions of being in public-facing positions. Many women are apprehensive about having a public presence online and expressing any public opinion for fear that they or their family may be targeted by abusive behavior [27].

One may argue that content moderation systems are not to blame for women disengaging from political discourse spaces online. This is partially true. Moderation systems are not solely to blame. Other platform practices and affordances, such as anonymity, the virality of posts, and the potential of coordinated networked attacks, all contribute to increased mis/disinformation and toxic behaviors. Political polarization and the increase of hateful public rhetoric spewed by politicians definitely hold partial blame. The continued existence of automated bots also holds partial blame.

Although the work of this chapter highlights violent and threatening speech as a way to illustrate how the 'US-interpretation of free speech' shape social media policies and platforms, other speech-related harms, such as misinformation and hate speech, that shape and affect engagement and interpretations of free speech but have not been discussed in-depth. For instance, as discussed in Chapter 3, the widespread existence of political mis/disinformation shaped young adult Gen Z women's skepticism in social media being used as a tool for effective political change, but also as a space of discourse. Additionally, other factors such as one's self-image, prior experience, race, socioeconomic status, background, context, level of education, etc., all affect how they may choose to engage in a public space. While the work of this chapter is unable to comment on how these dimensions of one's identity, and other speech-related harms affect young adult Gen Z women's engagement in political discourse, it is important to acknowledge how one's experience is not just shaped by a moderation system or a platform's interpretation of free speech.

However, the purpose of content moderation systems is to mitigate behavior to enable a public space of discourse in which users feel that they may freely express their opinions. At present, it is not doing that. A significant number of users on these platforms do not feel that they can engage in free speech, despite the platform promising them they can. An

inadequate moderation system, rooted in a US-interpretation of free speech that centers a speaker and disregards an audience, may be contributing to this broader disengagement. In the following chapter, I explore an alternative interpretation of free speech that is rooted in broadly understanding speech as a relationship between the speaker and the audience. I examine how employing this interpretation in a social media platform may affect practices of engagement among young adult Gen Z women.

CHAPTER 5
RELATIONAL INTERPRETATIONS OF FREE SPEECH

5.1 Introduction to Chapter 5.

In this chapter, I build off of the work done in Chapters 3 and 4 to explore the potential of engaging a relational theory of free speech in social media spaces of political discourse. This chapter investigates **(RQ3) How might engaging a relational theory of free speech in social media speech moderation policies affect engagement among young adult Gen Z women in online spaces of political discourse?**

To examine this question, I first explore existing work on alternative interpretations of free speech, specifically relational theories of free speech. I draw on feminist legal scholarship, philosophy of language, computer-mediated communication research, and political science research to understand current frameworks of relationality in the context of free speech in the U.S. I then engage these frameworks for relational free speech in an ideative workshop study with young adult Gen Z women. In these workshop sessions, participants were encouraged to re-conceptualize speech moderation practices rooted in a relational theory of free speech. I find that participants advocate for moderation systems to incorporate mechanisms for negotiation, reflection, and human intervention. I also find that while participants are appreciative of moderation systems being more situated and relational under a relational theory of free speech, they feel uncomfortable with platforms having access to personal information about users.

I then discuss how participants of this study perceived relational theories of speech and how this contrasted with their existing notions of what is 'free speech'. I argue that the culturally pervasive US-interpretation of 'free speech' inhibited participants' ability to challenge this interpretation and think of a moderation system that did not function

within this interpretation, in addition to how they might engage in public political discourse that does not operate under this interpretation. I broadly argue that in this case, the US-interpretation of free speech restricts one's ability to question whether they are enabled to engage in 'free speech' in public political discourse.

5.2 Theoretical Background - Relational Theories of Free Speech

In the United States, notions of free speech rooted in free speech absolutism and Libertarian philosophies are widespread, influencing law, policy, and the Internet (See Chapters 2 & 4).

5.2.1 Reconciliation of Values in Freedom of Expression

Many legal scholars have raised concerns regarding the relevance of these theories of free speech in the 21st century [22, 21, 23]. How might Mill's truth model, developed in the 1800s, still be applicable in a time where mis/disinformation runs rampant online? Kai Riemer and Sandra Peter argue that current platforms need to reframe their understanding of free speech, specifically in the context of content moderation. They argue that the algorithmic dissemination of speech on platforms forces users to become an audience for speech they may not have originally wanted to engage with [240]. Legal scholar Tim Wu argues that in the present day, free speech is a tool used to suppress other speech. Legal scholars Toni Massaro and Helen Norton build on this argument, explaining how the 'availability' or 'access' to tools that enable widespread speech, such as social media, also enables those who want to stoke division and distrust, exacerbate fears, and manipulate public opinion, which can negatively impact democratic institutions. For instance, disinformation and malinformation are considered mechanisms to undermine election processes by the U.S. Cybersecurity & Infrastructure Security Agency (CISA). CISA, a department under the broader Department of Homeland Security, has invested a significant amount of resources in battling disinformation during election years [241, 242], indicating how big of

a threat governmental departments consider disinformation and 'false speech'.

Massaro and Norton further argue that the volume of public speech that occurs makes it difficult for individuals looking at it to decipher what is true and what is not [22]. This is at odds with Mill's truth theory, which explains that the truth is 'discoverable' and will rise out of debate. Massaro and Norton argue that Mill's theory was acceptable in an arena where fewer voices were competing to shape public opinion, however in present-day, it is difficult to decipher the truth when one is inundated with information, and often, this information is viral and anonymized [7, 20]. Massaro and Norton's argument is also at odds with the autonomy theory of free speech which advocates for minimal regulation on any form of speech [76]. Many advocates for free speech absolutism call for counterspeech as a solution to mitigate mis/disinformation [41]. However, Massaro and Norton argue that counterspeech is not necessarily a feasible solution in environments with large volumes of public speech, such as on social media, as it may become even more difficult to distinguish what is false speech and what is truthful speech [22]. Legal scholar Michael Glennon advocates that 'more speech' is the solution to address mis/disinformation, rather than moderation mechanisms, but amends his claim to state that it is difficult to address deepfake AI-generated images through counterspeech [243]. Furthermore, Glennon's statement regarding the difficulty of countering AI-generated content raises questions regarding how effective counterspeech is against automated bots creating speech. Much of the disinformation and malinformation that occurs online is perpetuated through automated bots that have the ability to continuously post information at an incredibly fast rate. The sheer volume of potential false content generated would be difficult to counter with counterspeech. Additionally, even if counterspeech were created to mitigate the false speech, it would be difficult for users seeking information to determine what is true and what is false. While platforms have mechanisms in place to mitigate the existence of these bots [244], if they are able to spread a piece of disinformation before being removed, the harm may already be done. Using counterspeech to address falsehoods perpetuated by an automated tool

designed to spread disinformation in a viral manner may not be the most effective solution.

Many scholars have also argued that the theories of free speech highlight individual values such as autonomy, self-realization, and dignity, but the implementation of these theories requires reconciling these values with other ones [245, 246]. In his book The System of Freedom of Expression, First Amendment scholar Thomas Emerson states, "Any system of freedom of expression must also embody principles through which exercise of these rights by one person or group may be reconciled with equal opportunity for other persons or groups to enjoy them. At the same time, the rights of all in freedom of expression must be reconciled with other individual and social interests. It is this process of reconciliation that has given rise to most of the controversial issues in the past and continues to be the major focus of attention today" [247]. In other words, Emerson puts forth that a system or theory of free speech should not negate other values in order to bolster the value of free speech and should not encourage the free speech of one over the free speech of another. The lack of reconciliation of these values can be seen through both Supreme Court Justice Roberts and Gorsuch's Opinions in which they argue that we must continue to protect speech that can cause 'incalculable grief' (See Chapter 2). One may argue that the value of the speaker's autonomy triumphs over the value of the intended recipient's well-being. Or the value of a speaker's self-fulfillment overrules the value of the intended recipient's feelings of safety.

5.2.2 Relational Theories of Free Speech

One particular interpretation of free speech that has had limited exploration is the relational interpretation of free speech. Feminist theorists often engage relationality to closely examine concepts such as power and autonomy in larger institutional structures [248]. Within law and First Amendment theory, feminist scholars such as Susan Williams engage relationality to put forth alternative interpretations of 'free speech' in the United States. Williams argues that Mill's truth theory and the autonomy theory of free speech may be more effective if reframed under a relational view of truth and autonomy. Feminist legal scholars have

long argued that the First Amendment has a race, class, and sex attached to it, and therefore excludes many from its protections [66]. Williams' reframing of free speech aims to address this.

Williams's argues that the "truth theory" is rooted in Cartesianism which is inherently limited. Proposed by René Descartes, Cartesian thinking leads to the creation of dichotomies, such as culture/nature, objective/subjective, and mind/body. These dichotomies then lead to the assumption that there is such a thing as one truth, it is objective, and that it can be discovered. Thus, since Cartesian thinking creates and justifies gender differences, portraying them as objective truths and disregarding the idea that sociocultural context and values influence one's interpretation of the truth, feminist scholars believe the 'truth theory' is flawed and "morally objectionable" [76]. Williams argues that a relational model of truth views 'truth-seeking' as a process that is worth protecting due to the social relations developed between people in the process. In viewing this as a process and as a relationship, a relational model of truth 'reintegrates the ethical and political values that a more Cartesian model of truth attempts to exile from epistemology' [249, 250].

Susan Williams further argues that engaging a relational perspective of 'autonomy' would also reframe the autonomy theory more effectively as well. She explains that autonomy as a concept cannot be separated from the relational, as to understand how to be autonomous, you need to understand communal relations. However, existing notions of autonomy with regard to free speech imply that speech is a mechanism by which one becomes autonomous and separate from the community. Williams explains that this 'model of autonomy...does not give sufficient weight to the role of communal norms in grounding and facilitating autonomy (and vice versa)' [249]. Williams does not create a new theory of free speech but rather reframes the existing one pervasive in the US to reflect relationality. In doing so, she argues that this interpretation of the truth theory and the autonomy theory may be more inclusive and enable more free speech among those who are often pushed out of public discourse.

Other scholars have also engaged relationality as a concept to reframe communication in the context of 'free speech'. Philosopher Jovy Chan makes an argument to view free speech as a 'two-way-right', as the culturally pervasive interpretation of free speech do not consider the relationship between the speaker and the audience and how it influences their communication [251]. Chan argues that Mill's theory and the autonomy theory exclude understandings of speaker/audience relationships in their respective perspectives on free speech and justifications for why speech should be protected. In doing so, they hide 'the true relational and interactional feature that is the essence of communication'. Chan puts forth the 'Two-Way Approach' interpretation of free speech in which speech is considered 'free' when 1) one has the right to 'choose who they communicate with, 2) one has the right to refuse to enter into particular communications, and 3) one has the right to "explore and invite willing partners to enter into communicative relationships" [251]. Furthermore, Chan's 'two-way-right' approach highlights the continuously changing roles in a communicative relationship. A speaker becomes the audience and vice versa as communication occurs, showing how these roles of 'speaker' and 'audience' within a relationship are dynamic.

Chan's interpretation of free speech shares similarities with Feminist language philosopher Jennifer Hornsby's view of communication. Hornsby critiques free speech absolutism by arguing that their interpretation of speech fundamentally misunderstands how speech occurs between parties. She puts forth the idea of 'reciprocity', arguing that it is a mechanism by which a hearer understands the speaker's intention during speech. Hornsby utilizes the context of hate speech to illustrate how "reciprocity" differs from the pervasive understanding of speech. Hornsby states that the US-interpretation is rooted in the mental intermediation doctrine, in which the hearer's own interpretation of the words the speaker said is rooted in the hearer's psychology. For instance, if Speaker A were to say something and Speaker B was offended, it is because of the way Speaker B understood Speaker A's words. Hornsby further explains that this point is rooted in a language philosophy in

which speech is offensive only if the hearer believes that the speaker intended for them to believe that the speech was offensive. Hornsby argues that this is unrealistic about the hearer's role in speech. She argues that hearers do not sit there trying to figure out what the speaker intends them to believe. Through reciprocity, hearers of speech simply do the task of understanding the words that a speaker is saying and a speaker is in charge of what those words are. She states, "Hearers share the thoughts of those who speak to them. And insofar as they share a language with those who speak to them, it is well nigh impossible to avoid sharing thoughts they hear expressed" [211]. Hornsby's theory of reciprocity is rooted in the relationship between the speaker and the hearer and fundamentally challenges the culturally pervasive free speech understanding that speech is just words that are said from one party to another (See Chapters 3 & 4).

Political Science scholars Matteo Bonotti and Jonathan Seglow also closely examine speech from a relational lens. They put forth a relational view of free speech that has three primary dimensions: permissibility, autonomy, and answerability. Permissibility is a speaker's liberty to speak in the first place, autonomy is a speaker's ownership of the words that they say, and answerability is a speaker's accountability for the words that they say and the consequences of those words. Under this relational view, speech is 'a form of recognition; at the highest level of generality, recognition of each other's status as free, autonomous and answerable participants in a shared social world'. Bonotti and Seglow argue that speech that denies an individual these dimensions, such as hate speech, is not free speech. They argue that hate speech denies the targeted individual permission to speak, denies that individual the ability to autonomously develop their perspective, and does not recognize that they are 'answerable to others for their speech' [252].

All of these relational interpretations of free speech center the relationship between the speaker and the audience of the speech, in addition to accounting for the situated context in which the speech was made, to whom it was made, and how it was made. Through a workshop study with young adult Gen Z women, I engage these relational theories of free

speech in the context of moderating online spaces of political discourse to understand how this change may affect forms of engagement in these spaces. To do this, I present these relational theories of free speech to a group of young adult Gen Z women in the form of a workshop scenario and ask participants to ideate on potential forms of content moderation given this alternative understanding of free speech. In the workshop scenario, I encourage participants to center the relationship between speaker and audience, a key feature of relational theories of free speech, as part of their redesign of moderation practices. I elaborate on the structure of the workshop below.

5.3 Methods

In this study, I conducted two 90-minute in-person workshops with a group of 9 Gen Z young adult women. All participants identified as women, were born between 1997 and 2005 (18-26 years of age), and resided in Atlanta. The first workshop (Workshop A) took place on October 30, 2024, and had 5 participants. The second workshop (Workshop B) took place on November 7th, 2024, and had 4 participants. Each workshop session was audio and video recorded and took place on Georgia Tech's campus. Two observers were present in each workshop to take notes.

5.3.1 Recruitment

Participants were recruited through word-of-mouth and snowball sampling. Recruitment messages were distributed in one graduate-level course, and two undergraduate-level courses at Georgia Tech. The recruitment criteria for this study were that participants had to identify as women, be born between 1997 and 2005 (member of Gen Z and of voting age in the US), be able to participate in a 90-minute, in-person workshop on Georgia Tech's campus, and engage in any capacity on social media platforms, preferably Instagram. Of the 10 participants recruited, 9 signed up to participate in a workshop (See Table 5.1). I focused on Instagram during the workshop due to the popularity of the platform among young adult

Gen Z women in the previous two studies conducted (See Chapter 3).

After participants signed up for the workshop study, I sent them a short survey requesting demographic details such as age, race/ethnicity, and tenure on social media. I also asked participants about their past experiences with political discourse on Instagram or other social media platforms, in addition to their existing knowledge of social media content moderation systems. Of the 9 workshop participants, 4 indicated that they had no knowledge of how content moderation systems functioned.

I did not ask participants about their political affiliation or political activity online or offline. However, since both workshops occurred around the 2024 U.S. Presidential election, many participants expressed their political opinions and discussed their general political activity during the workshop.

5.3.2 Workshop Structure

Each workshop began with participants introducing themselves to one another and providing background on their general social media usage and previous experience with political discourse online. This was followed by a short presentation by the moderator of the workshop on how social media moderation systems typically work. The presentation included information about the various forms of moderation and a public example of when Twitter's moderation system failed. The example provided was of Democratic political strategist Rochelle Ritchie's experience with targeted threatening tweets. The individual making tweets threatening her was later arrested for mailing bombs to various high-profile politicians including former US President Barack Obama [253]. Twitter later apologized to Ritchie for dismissing her multiple reports of targeted threatening behavior [254].

Participants were also given paper copies of Instagram's community guidelines (as of October 2024) and an example political discourse post found on Instagram. Following the brief presentation, participants were provided with a speculative scenario in which they were asked to redesign Instagram's moderation system given two constraints.

The speculative scenario provided to participants was as stated:

> You are being recruited by Instagram to redesign their content moderation system. You are allowed to be as creative and out of the box as you would like because money and feasibility are not a problem here.

The first constraint provided to participants was that **the moderation system must take into account the relationship between the author of the post and the target of the post when taking any action**. The second constraint was that **the moderation system make must take into account the relationship's context and background when taking any action**. No further information to define these constraints was provided to participants. I wanted to understand how participants defined 'the relationship between the author of the post and target of the post' and how they defined 'speaker context and background'.

The two constraints posed to participants are derived from relational interpretations of free speech (Section 5.2.2), such as those that emphasize contextual and situated understandings of the speaker and their relationship to the listener/audience [252, 249, 251]. Participants were allotted fifteen minutes to individually ideate on redesigning Instagram's moderation system given the constraints and then the remaining time to share and discuss their ideas with the larger group. During the discussion portion of the workshop, all participants were first asked to share their ideas, which was then followed by a discussion on each idea.

5.3.3 Data Analysis

Immediately following each workshop, I met with both observers to debrief. The debrief sessions were 30 minutes in length and were audio recorded with the observers' permission. During the session, we discussed points of interest that emerged from the workshops.

Following these debrief sessions, I re-watched the video recording of each workshop session and used the notes from the debrief sessions as a starting point to iteratively code the video. The themes that emerged from this iterative coding were: 'Negotiation', 'Feelings of

Table 5.1: Participants in Workshop Study

P#	Workshop A or B	Age	Race / Ethnicity	Social Media Tenure
P1	A	23	Latina Mestiza (mixed)	9 years
P2	A	22	Chinese Indonesian	9 years
P3	A	24	Asian	6 years
P4	A	25	Asian	13 years
P5	A	26	Asian (Chinese)	9 years
P6	B	22	Mixed (Black/Asian)	12 years
P7	B	23	Asian	1 year
P8	B	26	Asian, Chinese Japanese	12 years
P9	B	21	Guyanese/Spaniard	9 years

Distrust', 'Desire for Human Intervention', 'Disengagement', 'Transparency', 'Accepting the limitations of the platform', 'Slow/Deliberate Engagement', and 'Avoidance'.

5.4 Findings

The findings from this study indicate that although participants felt that engaging in an alternative understanding of free speech within a platform may alter approaches to moderation, it would largely not affect their personal forms of engagement within online spaces of discourse. Participants were skeptical of how much an alternative approach to moderation may impact their own usage of a platform. Despite this, many common themes emerged from the ideas that were shared among participants during both workshop sessions.

5.4.1 Changes to Moderation Systems

Negotiation

Participants spoke of a desire to negotiate and or discuss the reasons why a post should be moderated. P3 in Workshop A spoke of wanting to not only know more about the

reasoning for a moderation decision being made but also the opportunity to provide their own reasoning and or motivations to mitigate a post online. Similarly, in Workshop B, P8 shared an idea that mirrored a 'customer service' approach where social media users could directly speak with an appointed representative regarding a moderation issue.

> "I was thinking like having- within the app, customer service type of thing where you're chatting with someone who is certifying... and having to navigate the conversation of a potential violation with a certified person and then having that conversation be reviewed later on by another non-biased person to like give a yes or no based on whether the post should be suspended" (P8 - Workshop B)

Other participants in Workshop B supported this idea, discussing how such a feature might also improve a user's relationship with the platform as it provided more clarity regarding why a moderation decision was made.

Throughout both workshops, participants spoke of wanting clarity and transparency on moderation decisions. They spoke of moderation inconsistencies they've observed through their own social media usage and felt that community guidelines and policies were opaque in their explanations. For instance, P1 in Workshop A shared a personal anecdote of an account of hers being banned on Instagram. P1 explained that there was no reasoning or explanation for why the account was banned, leaving her to speculate about what she may have done wrong. When asked about why they would like more clarity on the Instagram community guidelines she said,

> "...I would like to have a more clear definition of what Instagram for example means by "we believe" or "when we consider that". Specifically mentioning what they're looking at when making this statement, because I can see that they say like, oh, if we believe that there's a risk of harm. But how are you making

> this decision? Are you looking at language? . . . there's a lot of things here that like are open to cultural nuances, that I may not get. . . " (P1 - Workshop A)

Many participants also suggested ideas in which they could negotiate with the platform on the type of content they are exposed to. P7 in Workshop B shared an idea of adding a feature in which users may negotiate with the platform to opt out of seeing a specific type of content. P7 further suggested having an A-F rating scale, in which users may assign certain types of posts a score, and then request the platform not to show them content that is in a specific score category.

Participants' desire to negotiate with the platform broadly indicates their interest in 1) knowing more about how platform moderation decisions are made, and 2) being a part of the platform's decision-making process. Participants wanted to share their opinions with the platform and be provided the opportunity to discuss what the best mode of action was for a specific post.

Reflection & Reflexivity

Through their discussions, participants repeatedly mentioned moderation strategies that encouraged reflection and reflexivity among users before they posted online. Some participants shared ideas for strategies that would make users consider the effects of their post before they make the post. For example, P1 in Workshop A shared an idea for an automated mechanism that requested users to consider how their post may affect other users. This mechanism would then continue to question the user before they posted, forcing the user to reflect upon the potential larger effects of making a post.

> "So like my idea is before the user posts it detects that you're talking about this topic. . . and asks, what are you posting about? Like, how would this affect the subject. . . I feel like sometimes what happens is that people are not aware of what they're posting. Sometimes they are. So I was thinking of it as like,

could there also be a way to get the user themselves to be more mindful of what they're posting and why they're posting it?..." (P1 - Workshop A)

P1's idea indicates a desire for more reflection amongst users as to how others within their online community may be impacted by their social media posts. Furthermore, her comment indicates a desire for users to be more deliberate in their posts. P7 in Workshop B shared a very similar idea that advocated for self-reflection before a user posted. P7 suggested that posts go through an automated screening process in which users would be notified (via popup message), before posting, whether their content may cause harm to another user. P7 did not further elaborate on how they defined 'harm' in this case.

P2 in Workshop A shared an idea that similarly advocated for users to be more deliberate with their posts by adding more steps to the posting process. P2 discussed a moderation mechanism that would discourage users from making inflammatory statements by making the posting process more extensive. Users would be required to go through a mandatory orientation when signing up to join the platform, and through this orientation, they would be educated on the kinds of posts they would be allowed to make. P2's reasoning indicated that although such a mechanism would not necessarily prevent all forms of toxic or negative behaviors on the platform, it may dissuade many who would be making inflammatory comments from joining the online space in the first place.

" I was thinking like what if before you sign up, you have to go through a mandatory orientation or workshop? And I think it does two things. One, like people who are genuinely careful and care about this sort of thing would understand how they might not accidentally violate community rules. And at the same time, people who are too impatient... would be put off earlier... So it's sort of like, are you sure you want to do this? So, it's not like we can't stop them. If you're determined enough, it won't stop you, but it will, it hopefully might like put off people..." (P2 - Workshop A)

Both P3 and P4 in Workshop A were vocal in their support for this idea. However, despite their support, both P2 and P3 spoke about how even if such a mechanism were implemented, they would still feel uncomfortable posting their opinions in political discourse spaces online, indicating a deep distrust of the system. P2 further spoke about never feeling comfortable on social media, so despite the potential promise of a space with less inflammatory content, she would remain wary.

> "I am probably not the best person to talk about social media, because personally I have never been fully comfortable using social media because for me I'm constantly paranoid about like just other people seeing my post, the fact that it could be up there forever, even after I delete that, someone probably like took a screenshot of it and it's like sort of that paranoia that keeps me offline. I guess now that you mentioned it, yeah. Um all of my ideas have been like very aggressively kicking people off the internet" (P2 - Workshop A).

Trust in System

During both workshops, participants indicated a desire for more humans to be a part of the moderation process which may therefore increase more trust in the overall system. P4 in Workshop A shared an idea regarding community forms of moderation in which other users take a more active approach to report and moderate specific behaviors they see in discourse spaces. P4 suggested that users could choose to opt-in to the system and be anonymous when reporting which may aid in incorporating more user perspectives into the moderation process. While participants in Workshop A largely supported this idea, P2 and P3 did raise concerns regarding the potential for the mechanism to be abused by those who repeatedly report an individual for making a post they personally dislike. The disagreement in the discussion indicates that participants, while supportive of more human intervention in the moderation system, were wary of the potential bias that could occur when doing so.

Conversely, P8 in Workshop B discussed how knowing that more humans were a part

of the moderation approach reassured her that action would be taken.

> "I think it'd be cool to be able to report that and know that something's being done about it or like a real person is handling the situation instead of just like an automated whatever." (P8 - Workshop B)

P9 supported P8's point about having more human intervention in moderation approaches, further discussing how she would prefer more than one set of eyes looking at posts that were moderated. P9 felt that having multiple human moderators assess a specific post made her not only feel reassured that there was a thorough process to moderation, but also that it increased her general trust in the system.

> "It really just comes down to caution. I mean it's severe. It's a threat. Like the result of them overlooking a tweet became bombs going to multiple high profile people's doorsteps. And then that launches a whole different relationship with people on social media. Now it's sort of like, okay, well now when I turn in your report, I really can't trust you because, it was overlooked. . . People sometimes, they beat around the bush and that's something that can't be detected by any programming kind of thing. Well AI is improving, but like I said, not everything is always so clean-cut or specific." (P9 - Workshop B)

Participants in Workshop B also spoke of developing more robust training measures for moderators. P6 in Workshop B shared an idea for implementing a certification process for moderators. Such a process may make the training for moderators more robust and rigorous, and therefore increase trust amongst users that the online spaces will be adequately looked over and maintained.

> "So I think like depending on what kind of company, it's like Instagram or Twitter, you have specific specialized training to be qualified to be a content

> moderator. I'm thinking like some kind of certificate or some kind of assessment to say like, Hey, you're clear, you have done this before. You did some practice." (P6 - Workshop B)

P9 supported P6's idea about having a certification, further discussing that it would be important in this process to educate moderators on previous scenarios where moderation occurred incorrectly.

However, when asked about whether, if such a process was implemented, it would change P6's usage or understanding of social media spaces, P6 was unsure. She mentioned that knowing that content moderators had a specific level of experience and were qualified led to some degree of reassurance, but not necessarily any changes to her engagement.

> "Like our content moderators are certified, and we only use certified people with like more than three years of experience or I don't know, like have more than three certifications. Like they're qualified and they know what they're looking for. They've practiced, they've been tested for what's right and wrong. So that makes me more reassured on Instagram that I'm being protected." (P6 - Workshop B)

Desire for Context

Many participants across both Workshops A and B expressed the potential for users to share more of their background and context when in public spaces of discourse. However, while the participants in Workshop A appeared to think positively about the idea of users revealing more about their background and context within online spaces of discourse, participants of Workshop B were more hesitant. P1 in Workshop A shared an idea of platforms gathering more information on users who may be more at risk of being harassed on the platform such as those from marginalized communities, public figures, or activists. P1 felt that if moderation systems were to allocate more resources to monitoring accounts that

were interacting with these at-risk accounts, they could more adequately deal with toxicity targeting these accounts.

Participants also discussed how users sharing more of their own personal background may impact moderation approaches. P3 in Workshop A talked about implementing a policy for users to share their political affiliation when making political posts online. By doing so, users would have more information about where the poster was coming from when making a politically oriented social media post. In a similar vein, P5 shared an idea of the moderation system curating content based on attributes of one's identity. For instance, should the moderation system know that a user identifies as a woman, it would purposefully filter out content that is prejudiced against women. In doing so, the moderation system would ensure that users are not shown inflammatory content that targets aspects of their identity without necessarily censoring other users.

In contrast, the participants of Workshop B discussed feeling uncomfortable at the thought of platforms knowing information about their background and about their relationships with other users. P8 expressed hesitation in allowing a platform to know information about them and their relationships with others on the platform. P8 felt that in doing so, the platform may be forming a judgment of the user that may be inaccurate, and that it would go beyond the scope of what a platform should be doing.

> "Because that means that the moderation system is forming some judgment, right? like on me and the user and it might be assuming the wrong things that might not be accurate... But like, would I want the moderation system to know that I might be having some sort of drama with someone else? I feel like that goes beyond the service of the platform. I feel that the app will just turn into another thing, you know, like instead of like the mission of it, it is just more of like a therapist..." (P8 - Workshop B)

P8's comments indicate privacy concerns, and more broadly, a willingness to forgo moderation in order to maintain their privacy regarding their relationships with other users

on the platform.

P6 supported P8's hesitation, expressing concerns about whether this feature would be abused. P7 felt that it is unnecessary for a platform to know details regarding the relationship between two individuals, as in those situations, if they were being 'targeted' by someone they knew, they would just discuss it amongst their friends and keep it to themselves. Similar to P8, P7 indicated that she would prefer maintaining her privacy rather than having a more context-aware moderation system.

The difference in opinions between the participants of Workshop A and Workshop B was surprising, considering how similar many of their ideas were for moderation systems. This suggests that there are differences in how much one's ability to engage in free expression is valued. Drawing on legal scholar Thomas Emerson's work on the reconciliation of values in free speech [247], these participants faced a need to reconcile certain values in order to design a moderation system that did not restrict the speech of one group over the other.

In Workshop B, participants implied that they were willing to remain in a toxic online environment so that aspects of their personal information and relationships remained private. The participants of this workshop may have felt that they were compromising their ability to speak freely in public spaces of discourse to enable their right to privacy. In Workshop A, participants did not mention issues of privacy, instead advocating for platforms to know and access a significant amount of personal information regarding a user's identity.

5.4.2 Skepticism of Potential Change

Although participants were vocal in their support for moderation systems under a relational theory of free speech becoming more rigorous in mitigating online toxicity and false information, many remained skeptical of how much the public space of political discourse would change. When asked about how their online behavior would change in this alternative scenario, most participants said that it would not change. For instance, P2 (as aforementioned

in Section 5.4.1) discussed her general distrust of social media platforms which affected her ability to trust whether a platform rooted in an alternative theory of free speech would change her perspective. Other participants shared similar feelings of distrust and apathy towards platforms and their existing practices.

Many of these participants were long-time users of social media and their prior experiences with political discourse shaped how they engage in these spaces. Their experiences and forms of engagement were very similar to those shared among participants in previous studies I conducted (See Chapter 3). Many spoke of disengaging and or avoiding public forms of engagement in political spaces. Many also spoke of a fear of facing emotional, physical, and or professional ramifications for expressing their opinions in public spaces of discourse. For instance, P3 in Workshop A stated that they typically do not engage in any spaces or make comments online when they know it may invite 'hate'.

> I do not engage in anything I know will get hate online. By anything that I am posting be it, you know, like a proper post or be it a comment... like I will not do it in a space I know where I don't want people seeing me or I don't want people even showing any potential hate. So anything that can be considered in any way, I just don't engage. So I feel like it, it doesn't apply to me as an author online or as like a speaker online that I need to go through all of this just because I just don't do it. " (P3 - Workshop A)

When asked about how potential changes to the content moderation systems would affect their behavior online, P3 felt that it would not affect them because they do not engage in public forms of discourse in the first place.

Many participants also felt that changes to a moderation system did not address the 'real' problem they felt was causing such online toxicity in political discourse spaces. P2 in Workshop A felt that the 'real' problem was other users who engaged in toxic behavior online. Other participants in Workshops A and B felt that the 'real problem' was general political polarization which led to toxic behavior in political discourse. Many of these

participants felt that changes to a content moderation system, rooted in a relational theory of free speech, would do little to address either of these problems. They perceived that changes to a moderation system would temporarily remove or mitigate toxic behavior, but would do little to completely stop this behavior.

From the workshop sessions, it is clear that participants were deeply distrustful of platforms. They were frustrated by the inconsistency of existing moderation practices and the lack of transparency from platforms regarding how moderation decisions were made. Their desire for mechanisms to negotiate and reflect indicated a broader desire to be more directly involved in the processes that were affecting their engagement in online spaces. However, participants did not feel that their ability to engage in free speech was being compromised through platform practices and policies. They spoke of disengaging and avoidance, but did not speak of free speech. One may speculate that their long-time tenure and exposure to political discourse on social media have shaped their perspective on what 'free speech' is on these platforms. One may speculate that their previous experiences have normalized their self-censorship, and therefore, they may consider this within the realm of 'free speech'.

5.5 Discussion

5.5.1 Practices of Avoidance

Participants of both workshop sessions shared ideas for moderation systems that encourage avoiding conflict altogether, rather than mitigating existing conflict and toxicity. For example, P1, P2, and P7 shared ideas to deter users from making inflammatory comments by making the process of posting more extensive and tedious for users. Similarly, P5's idea to curate one's feed so that they may avoid posts containing content they dislike or find toxic enables users to avoid toxicity rather than combat it.

These ideas not only highlight a desire to avoid being exposed to content that is toxic and or controversial but also a desire to dissuade users from posting toxic or controversial content without explicitly forbidding them. During the workshop and my subsequent analysis, I found that participants preferred subtle mechanisms that "dissuaded" users from posting inflammatory content but that did not explicitly forbid them from doing so. For instance, P2's idea suggests that by making the process of posting more complicated and tedious, users would choose not to post inflammatory content without explicitly being told not to do so. Such ideas may further indicate that participants in these workshops did not necessarily break out of the pervasive US-interpretation of free speech, which dissuades any form of censorship.

Prior work in HCI and Design has highlighted the difficulties of asking study participants to challenge dominant theories and conceptions [255]. For example, a 2021 study using design-futuring methods exploring the ideas of youth in the US on their usage of technology in their future found that it was challenging for them to imagine completely new perspectives. The authors found that dimensions of oppression, such as racism and poverty affected participants' ideas, as most of the design-futuring ideas shared included the existence of racism and or poverty [256]. In other words, despite speculative and design-futuring methods encouraging participants to ideate without constraints, it is often difficult for participants to imagine worlds or situations without those constraints.

In the case of this study, participants had challenges re-imagining social media spaces rooted in alternative interpretations of free speech in which they would feel comfortable engaging. Their ability to re-imagine these spaces was influenced by two factors. One factor was their prior experiences in and knowledge of these spaces, which may have influenced their ability to completely re-imagine what these spaces may look like. Participants often referenced their personal experiences when explaining their redesign ideas, indicating that their experiential knowledge shaped their redesign. For example, P2 drew on her

experiential knowledge on Tumblr to inform her perspective on incorporating mechanisms for negotiation and reflection in new moderation practices. Additionally, participants often drew on the design and practices of existing platforms to inform their new designs. For instance, P8 drew on the design of existing customer service chatbots to inform their design of a moderator chatbot with which they could discuss and negotiate moderation decisions. The second factor was the widespread, pervasive understanding of free speech, which appeared to have influenced their ability to completely re-imagine the practice of moderating speech in public discourse. The ideas participants shared did not explicitly target certain forms of speech but rather enabled users to avoid speech that would negatively impact or offend them, indicative of their desire to avoid 'censorship'.

Perhaps the clearest indication that participants struggled with challenging the culturally pervasive interpretation of free speech is that they often only viewed themselves as 'speakers' when generating ideas for redesigning moderation systems. For instance, participants advocated for mechanisms to negotiate moderation decisions with the platform. Although their proposed mechanisms for negotiating could be applicable to a recipient of targeted speech, participants often placed themselves in the position of a speaker/poster wanting to get clarity on why their post was moderated. In another example, participants proposed mechanisms to encourage reflection and reflexivity among speakers/posters but did not propose any mechanisms to encourage reflection or reflexivity among recipients of speech. For instance, P1 proposed an idea that encouraged a poster to think through how their post might be interpreted before posting it, and P2 proposed an idea that forced a poster to be more deliberate with the language in their posts. None of the ideas proposed by participants encouraged the recipient of the speech to reflect on anything. Furthermore, none of the proposed mechanisms were for those who have been targeted by speech. Participants did not propose any ideas for how an 'intended audience' may address toxic speech targeting them.

The culturally pervasive US-interpretation of free speech that shapes existing moder-

ation practices centers the speaker and their needs over those of a recipient of speech. Participants appeared to draw on this interpretation of free speech by placing themselves in the role of a speaker when asked to redesign moderation practices, despite being asked to explicitly center the relationship between speaker and audience in these new moderation practices. However, participants' practices of avoidance on their own social media accounts indicate that they do not perceive themselves as 'speakers' when engaging on social media platforms. While some participants, such as P5 and P7 proposed ideas for moderation mechanisms that may encourage users who lurk to participate more publicly, many participants did not propose ideas for moderation practices to encourage speech among 'non-speakers' on platforms. Participants attributed their avoidance on platforms to the actions of speakers/posters, but did not perceive how the roles of speaker and audience were interchangeable nor that their avoidance may be evidence of a compromised ability to engage in free speech. The ideas they proposed addressed a 'speakers' ability to engage in free speech, but not necessarily the ability of a recipient or those who disengage, avoid, and or observe to engage in free speech.

Furthermore, their subsequent skepticism of potential change that could occur on platforms if a relational theory of free speech was incorporated into the design of moderation practices is further evidence of their struggle to question the dominant interpretation of free speech. Despite many participants explicitly discussing their role as 'lurker' or 'observer' in online spaces of discourse, almost all positioned themselves as 'speaker' when asked to redesign a moderation practice. Their skepticism appeared to stem from the role that they personally embody on their own social media, however, their proposed moderation mechanisms that center the needs of a 'speaker' appeared to stem from continuous exposure to the dominant interpretation of free speech. While these two perspectives are at odds with each other, it is evident that participants struggled to perceive their personal experiences online as evidence of a compromised ability to engage in free speech.

However, it is important to note the nuance in how participants viewed speech and free

speech. Although participants struggled with challenging certain dimensions of the culturally pervasive interpretation of free speech, they did challenge other aspects of it. The experiences of participants in this study were quite similar to those in Chapter 3. In Section 3.5.6, I found that participants of the interview and diary study viewed speech as fundamental to their relationships, as they were cautious to engage in speech that they felt would negatively affect their relationships. They viewed speech as beyond 'just words', contrasting greatly with the US-interpretation of speech. In that section, I raised the question of whether platforms share their interpretation of speech and free speech with their users. The findings of Chapter 3 appeared to indicate that 'no, platforms and their users do not share the same interpretation of speech'; however, the findings of this chapter seem to complicate these prior results. Participants of the study presented in this chapter seemingly abide by the culturally pervasive notion of free speech, sharing ideas for moderation practices that adhered to this pervasive interpretation. However, their ideas and individual forms of engagement online indicated that they too, perceived online speech as beyond 'just words'. For instance, their ideas that encouraged more reflection and reflexivity indicate an understanding that speech has an impact beyond 'just words' (Section 5.4.1). Although their perspective may have been 'speaker-centric', their focus on enabling reflection and reflexivity among speakers when posting indicates that they view speech as having an impact on the physical, emotional, and mental well-being of the intended audience. P1 and P2's ideas encouraged speakers to be more deliberate and empathetic before posting, thinking through how their post may be understood by their intended audience. This idea broadly indicates an understanding that speech is situated and is fundamental to shaping a relationship between two parties. When tasked with re-imagining moderation practices, participants both adhered to and subtly rejected the culturally pervasive, US-interpretation of free speech.

However, their struggle to completely challenge the US-interpretation of free speech when redesigning moderation practices serves as further evidence of how pervasive and dominant the interpretation is in both online and offline discourse.

5.5.2 Participant conceptions of relational

The findings from this study indicate that participants had varied perspectives on incorporating a relational interpretation of speech in content moderation systems. Several participants shared ideas for moderation mechanisms that supported the relational free speech theories presented in Section 5.2. For instance, P1 and P2's ideas encourage users to reflect and be 'answerable' for the online speech they make, as described in Bonotti and Seglow's interpretation of relational free speech [252]. Participants also shared ideas that enabled moderators and moderation systems to have more information about users and their situated contexts. For instance, P3 and P5's ideas to create mechanisms by which users may choose what content they prefer to interact with based on varying dimensions of their identity. This idea incorporates aspects of Chan's conception of free speech as a two-way right, specifically the right to choose who to engage with [251].

However, some participants were also apprehensive about platforms and moderation systems learning more information about users, as aforementioned in Section 5.4.1 (*Desire for Context*). P7, P8, and P9 in Workshop B expressed concerns for their privacy if the moderation system had access to more information about them and their relationships when mitigating toxic content. These participants appeared to consider the integration of a relational interpretation of free speech as detrimental to their privacy in these online spaces. As discussed in Section 5.4.1 (*Desire for Context*), some participants felt that engaging in a relational theory of free speech meant compromising other important rights. However, participants did not question whether their ability to engage in free speech was compromised by existing platform moderation policies. One may argue that perhaps, through their prior experiences in these public spaces of discourse, their compromised ability to engage in free speech is their only understanding of 'free speech'. In other words, their only understanding of free speech is a partial right to free speech.

Practices of Engagement & the Limitations of Moderation

Participants who supported incorporating the relational interpretations of free speech into moderation practices and those who were apprehensive of it indicated that regardless of the change, their personal patterns of engagement in these spaces would not change. For instance, P2 and P3 expressed that even with significant changes to the platform and moderation systems, they would be hesitant to be open about their opinions in online spaces of political discourse. Similarly, P6 in Workshop B felt unsure of whether her behavior in spaces of political discourse would change even if she knew that content moderators reviewing toxic content on the platform had a specific level of certification. These participants' apprehension is indicative of a distrust of platform moderation practices, but also more generally, indicate an understanding of the limitations of moderation practices.

This finding supports previous findings found in earlier interview studies (Chapter 3). In Chapter 3, I present findings from an interview study with young adult Gen Z women that drew attention to the skepticism and distrust participants had in social media platforms. The findings from this workshop study are similar, indicating that even with large, substantial changes made to the content moderation systems of these platforms, users feel uncomfortable engaging with discourse online. In this study, I found that participants were resigned to the fact that online spaces of public discourse were hostile, toxic, and argumentative. Prior research found that exposure to such toxic behavior online had led many women-identifying users to adjust their behavior online to account for or in anticipation of hostility [166, 203].

Similarly, the participants of this workshop study appeared to have accepted and resigned themselves to the idea that moderation practices online are limited and unable to address the various forms of toxic behaviors they encounter online. Participants were aware that moderating content is a challenging process and that, given the scale of social media, it is difficult to ensure that all toxic content is adequately moderated. It appears that their prior experiences and exposure to social media shaped their perception of what moderation

practices could do. Furthermore, as discussed in Section 5.4.2, participants felt that moderation would not address the 'real issue' which for many, was societal political polarization. From their perspective, any public political discourse space would be hostile in the current political climate, thus, changes made to moderation systems could not address this. Similar to the findings of the 2021 design-futuring study [256] discussed in Section 5.5.1, in which participants struggled to imagine futures without oppression or racism, participants of this work struggled to imagine engaging in political discourse without extreme political polarization.

5.5.3 Challenging Understandings of Free Speech

As noted in Section 5.5.1, it is evident that the pervasive US-interpretation of free speech restricts one's ability to challenge their own interpretation of free speech. Participant's unwillingness to break away from common notions of 'censorship' and existing forms of social media moderation policies indicate that their perspective continues to be largely shaped by the dominant interpretation of free speech. Furthermore, in the present day, rather than questioning such dominant interpretations, social media platforms have continued to reinforce them as well. In a statement released in January 2025, Meta announced that in a move to increase 'free expression' on their platform, they will be removing their formal fact-checking program [32]. They also announced that their automated systems, which previously were used to scour for violations of all content policies, will now only be used for high-severity violations such as terrorism, child sexual exploitation, and fraud. For less severe violations, such as hate speech, the platform will not take action until content is reported. Meta refers to this mode of moderation as the 'Community Notes model', in which the community leads efforts to maintain and mitigate their own space [32]. Some of these changes to Meta emulate what participants spoke about in this study, particularly the emphasis on user participation in these efforts. However, while participants spoke of more nuanced moderation mechanisms and reflection among users prior to posting, Meta's

changes are pushing for less moderation and more speech in general. While one may argue that Meta's push for less moderation and more speech is to encourage more usage and is potentially financially motivated, the company's statement attributes their changes to a desire to encourage 'more free expression'. Although Meta might be using 'free speech' as a disguise to hide their true motivations for changing their moderation policies and mechanisms, it is difficult to deny that these recent changes highlight the tight hold that free speech absolutism and the Libertarian models of free speech have over the United States.

The findings of this study and the recent continued reinforcement of free speech absolutism, draw attention to the difficulties of challenging these dominant perspectives in a platform context. How does one challenge these widespread notions of free speech that inhibit their ability to speak freely? Furthermore, is it possible to use one's right to 'free speech' to challenge the dominant interpretations of 'free speech' in a social media context?

5.6 Conclusion

In the following chapter, I engage design fiction methods as a thought exercise to challenge the traditional US-interpretation of free speech in online platform policies. I create a design fiction of a platform content moderation system rooted in relational theories of free speech and use it to question how we may challenge the US-interpretation of free speech in an online context.

CHAPTER 6
ENGAGING RELATIONAL THEORIES OF FREE SPEECH TO RE-CONCEPTUALIZE PLATFORM PRACTICES

The work presented in the previous chapter (Chapter 5) demonstrates that the continuous exposure to free speech absolutism through social media platforms may serve to restrict one's ability to question whether this interpretation of free speech is truly enabling free speech rather than limiting it. The findings of that work raised the question, how might we challenge this widespread interpretation of free speech? I use design fiction as a thought exercise to explore this question.

In this chapter, I present a design fiction of a content moderation system rooted in relational theories of free speech that also draws on findings from an ideative workshop study conducted with 9 young adult Gen Z women (Chapter 5). This design fiction is of a fictional content moderation system called *Agora*. Building on the themes of negotiation, human intervention, reflection/reflexivity, and concerns regarding context that emerged from the workshop study conducted with young adult Gen Z women, Agora aims to enable community-based content moderation practices rooted in relational theories of free speech.

Although I engage design fiction as a method, the contribution of this chapter is not the design itself. Instead, I engage design fiction as a thought exercise to illustrate the potential of drawing on relational theories of free speech in a platform content moderation practice and to explore how it may impact forms of engagement in spaces of political discourse online. Through Agora, I re-conceptualize moderation practices by rooting them in a relational theory of free speech. Furthermore, I engage Agora as an instantiation to highlight how the practice of re-conceptualizing may serve to question and challenge sociopolitical concepts, such as 'free speech', that are ingrained within U.S. culture.

6.1 Background

Design Fiction has been used as a method by researchers in the HCI and Design community to re-imagine and explore 'what could be' [257]. Rooted in the traditions of science fiction, design fiction integrates literary, speculative, and design practices as part of its methodology. Design fiction often interrogates concepts by reframing them under fictional contexts or redesigning them to expose underlying tensions [258, 259]. For instance, a 2023 study engaged design fiction as a tool to imagine a social media platform, The Stoop, designed by and for Black users. As part of the study, researchers created a fictional backstory for why and how the tool was designed and developed. The researchers also created a prototype of the tool and fictional promotional materials to demonstrate how users may interact with it. Part of engaging this design fiction method was the creation of an entire narrative around the fictional social media platform in order to fully re-imagine the potential of what a space like The Stoop might be [260].

In this chapter, I use design fiction as a tool to reveal how engaging a relational theory of free speech in a content moderation context may differ from existing content moderation practices. I create a fictional content moderation system rooted in a relational theory of free speech to illustrate what engaging a relational theory of free speech might look like.

6.2 AGORA

This section introduces the design fiction, Agora. The following subsections are within the fictional narrative surrounding Agora.

6.2.1 Description

"Agora": borrowed from Greek *agorá*, noun derivative of *ageírein* "to gather, assemble" [261]

Agora, launched in May 2027, is a community-based content moderation service that

has since been implemented into major social media platforms such as Facebook, Instagram, and TikTok. It works in tandem with existing platform moderation mechanisms. Agora operates as a service to moderate directly targeted posts online. The service aims to provide both parties of the online speech a space to engage in a dialogue regarding their interpretations of the content and its potential impact.

Background

Agora is designed in the style of an online forum. When an online post is flagged by the platform, the original poster and intended recipient may use Agora to discuss and negotiate how the post should be moderated, if at all. As mentioned earlier, only directly targeted posts can be evaluated through Agora. For instance, posts that directly tag an individual or directly mention another user's name can be evaluated through Agora. Posts aimed toward a group or reference a population of people will not be considered.

After the original post has been brought to Agora and both parties involved have provided context and perspective, 50 members of the Agora service will vote on whether and how the post should be moderated based on the discussion. Participation in Agora is optional for users. If both the intended recipient and original poster do not use Agora, their post will be evaluated using the existing platform moderation mechanism. If only one party opts to use Agora, the other party will be notified and will have the option to accept or decline participation. Should they not answer, the other party may continue to use the service but will be required to provide additional contextual information to compensate for the lack of one party's participation. Should they decline, then neither party may use the service, and the post will be moderated by the platform moderation rules.

As Agora is designed in the style of an online forum, the service has a set of moderators to ensure that community members feel that the service is meeting their moderation needs. These moderators are selected based on their prior participation and experience in the community. All moderators are required to go through training and are briefed on previous

instances of community infractions and how they were handled.

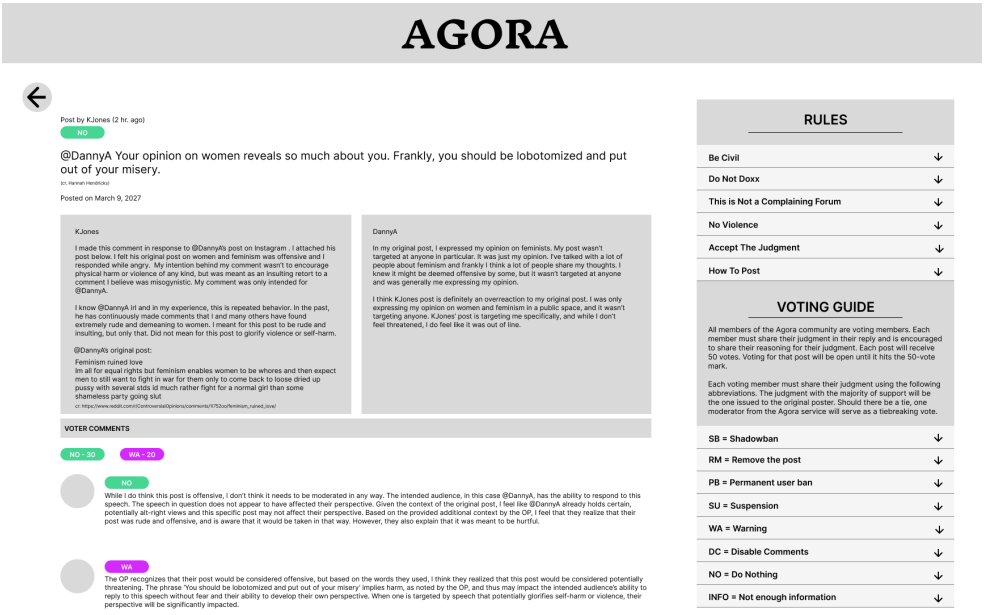

Figure 6.1: Members of Agora voting on and discussing the post

Usage

Users share the post in question with the larger Agora community to discuss whether it should be moderated and, if so, how. This post is required to anonymously share the following contextual information with the community: when the post was made, the post's intended audience and their reaction to it, and the original poster's intention behind the post. Both the original poster and the intended audience may share additional contextual information, such as information about the original poster's relationship with the intended audience, the original poster's past history with making certain posts, and whether the post was in response to another post. While sharing this additional information is encouraged, it is up to both the original poster and the intended audience as to how much they would like to share with the broader community. The purpose of this space is to give the original poster a chance to explain their perspective and reasoning for their post, the intended audience a chance to explain their perspective and reaction to the post, and a space to negotiate with

the broader community on how to best moderate the content, if necessary. When a post is being discussed in Agora, it will be taken down from its original platform. Following the judgment reached by the Agora community, the post will be appropriately handled by the platform.

Suppose only one party (either the original poster or the intended audience) is present when posting in Agora. In that case, the user bringing the post to the community will be required to share the following information: the original poster's relationship with the intended audience, the original poster's past history with making certain posts, and information about the reaction the post received when originally posted.

6.2.2 How To Post

Users must include the following when sharing a post with the Agora community:

- Are you the Original Poster, the Intended Audience, or are both present (Original Poster + Intended Audience)?
- When was the post made?
- Who was the post's intended audience, and what was their reaction?
- What was the original poster's intention behind the post?
- How did the original poster think the audience would interpret the post? What did they think their reaction would be?

Feel free to ANONYMOUSLY share any additional context that may help the community make their judgment.

In many instances, both the original poster and the intended audience participate in the discussion presented to the voters. If this is the case, select the 'Original Poster + Intended Audience' option when creating the post. When posting, please anonymize yourself and others you are discussing. As mentioned above, you may only reveal information for the

context that is generic and or vague. If you reveal information that is not anonymized or doxx another user in the process, you will be permanently banned from the Agora community, and your post will be removed from Agora.

Agora posts are limited to 700 words, so please only provide information that is relevant and specific to help the community make their judgment. If a user continuously makes posts that violate the specified guidelines, they will be served a warning and may be temporarily banned from the Agora community.

6.2.3 The Judgment Process

Agora adheres to a relational understanding of 'free speech'. Specifically, voters are encouraged to better understand the relational and interactional components of the online speech they are evaluating. For instance, how might both the original poster and the intended audience of the post interpret the content in question? Additionally, how might their relationship or interactions affect interpretations of the content?

All voters must consider three important points to determine whether the content in question is 'free speech' or not.

- In this situation, does the intended audience now have the opportunity to reply to this speech without fear of physical, emotional, and or financial retaliation?
- In this situation, does the intended audience now have the ability to develop their perspective autonomously?
- In this situation, does the original poster recognize that they are held accountable for their speech and the consequences of their speech?

All voters are asked to keep these three questions in mind to guide their judgment process.

6.2.4 Voting Guide

All members of the Agora community are voting members. To participate in a vote, the member must share their judgment using the options listed below. Each voter is also encouraged to share their reasoning for their judgment through a commented reply. Each post will receive 50 votes. Voting for that post will be open until it hits the 50-vote mark. The judgment with the majority of support will be the one issued to the original poster. Should there be no clear majority, a second round of voting among the members will be conducted. If there is a tie, one moderator from the Agora service will serve as a tie-breaking vote.

The list of following moderation options is consistently updated. Options may be added and or removed as time goes on. If a community member has a new option they would like to add to the list, they may provide a description of the option and an explanation for why it should be added to the moderators of Agora. Similarly, if a community member would like to remove an option, they may explain to the moderators why it should be removed. These moderators will then set up a community-wide vote to determine the addition or removal of a moderation option.

Voting Rules

SB = Shadowban

Shadowbanning occurs when a post is purposefully removed from certain spaces on a platform. To the poster, it appears that the post is still up and available, however, it will not appear across many other users' feeds and will receive significantly fewer impressions.

RM = Remove the post

Removal of a post means that the post will be permanently removed from the platform. No users, including the original poster, will be able to access or view the original post.

PB = Permanent User Ban

If a user is permanently banned, they will no longer be able to access their account. All of their existing posts will be automatically removed from the platform. They will also not

be able to create a new account using their existing credentials (username, email, phone number, etc)

SU = Suspension

If a user is suspended they will not be able to access their account for a specified period of time. Suspensions can last for 60 minutes, 1 day, or 1 week. The length of suspension is determined by the community and is dependent on the infraction.

WA = Warning

If a user is issued a warning, no action will be taken on their original post. However, this warning will be accounted for on the user's Agora profile, and may be factored into evaluations of any potential future posts they bring to Agora.

DC = Disable Comments

The original post will have its comments section disabled. No user will be able to comment on the post.

NO = Do Nothing

No action will be taken against the post or the poster.

INFO = Not enough information

If the community issues an 'INFO' judgment, additional information must be provided for the community to make a judgment. Community members should specify, in the reply section, what additional information is necessary for them to make their judgment.

6.2.5 The Rules of Agora

Below are the rules users must follow when participating in Agora.

Be Civil

We request all users in Agora be civil with one another. We encourage discussion and debate, however, we do not tolerate any personal attacks on other users. Do not make comments on another user's race, gender identity, sexual identity, or other such personal

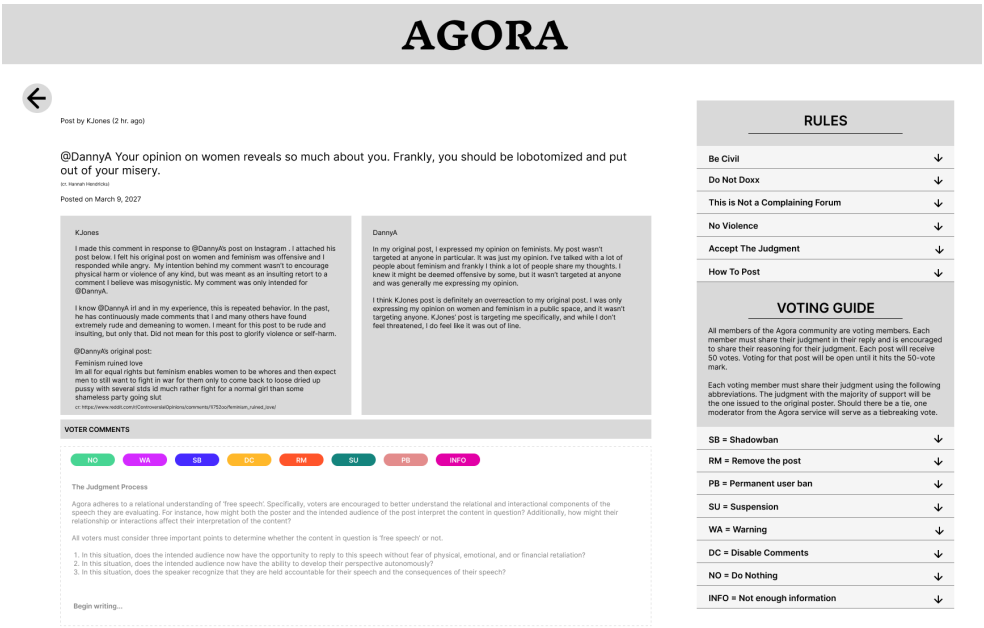

Figure 6.2: Voter view of discussion between original poster and intended audience

characteristics. If this occurs, you will be banned from the Agora community. If banned for this reason, you will no longer be eligible to join Agora again.

Do Not Doxx

The Agora community has a strong stance on doxxing. If you reveal another user's personal information, such as their name or address, you will be removed from the Agora community, and your post will be immediately deleted. When posting, you may reveal certain characteristics of yourself or others in your post IF and ONLY IF they are anonymized and or are generic. For example, you may reveal in your Agora post that you identify as a woman and are good friends with the intended audience of the post you are discussing. You may also reveal that you have a professional background in [blank], which informed your post. But do not reveal your name, your address, where you went to college, what your specific job role is etc. If you are unsure of whether the information you are sharing is not anonymous enough, do not share it and indicate in your post that if you reveal that information, it would compromise anonymity.

No Violence

The Agora community has no tolerance for violent and threatening speech. Should a user mention, wish for, or call for an act of violence or implicit harm (physical, emotional, & mental), they will be permanently banned from Agora.

This is Not a Complaining Forum

Oftentimes, when discussing a post, users delve into complaints about the social media platform, complaints about other users, etc. This is not a space to air out your grievances. The purpose of Agora is to provide a space for in-depth discussions on the content moderation of posts that have been flagged. Agora provides space for dialogue between the original poster and the intended audience regarding interpretations of an online post. If this dialogue begins to veer into users airing out their personal grievances with each other that have little to do with the post, we will remove this post from the discussion circulation and let the original platform handle the post in whichever way they see fit. Furthermore, members of the Agora community are not counselors or therapists and cannot help you resolve interpersonal conflict. Please remain on topic when making a post.

Accept The Judgment

Once the community has issued its judgment, accept it and move on. Please do not continue to debate or argue regarding the judgment. If this occurs, you will be removed from the Agora community. Should you feel that the community has made an error in their judgment, you may choose to decline the judgment, and your post will go back to the platform and be evaluated in whichever way they see fit.

6.3 Inspiration for Design

This section discusses the underlying theory and inspiration for the design of Agora. This section and onwards are no longer part of the fictional narrative surrounding Agora.

6.3.1 Engaging a relational theory of free speech

The findings of the workshop study (Chapter 5) indicate that engaging a relational theory of free speech in social media moderation policies does little to affect engagement in online political discourse spaces among young adult Gen Z women. The widespread US-interpretation of free speech may have shaped these participant perspectives of what is considered 'free' and what is not, influencing their ability to challenge the boundaries of this pervasive interpretation of 'free speech'. However, this design fiction, Agora, aims to challenge the pervasive interpretation of 'free speech' by presenting an alternative content moderation system rooted in relational theories of free speech. The purpose of Agora is to demonstrate the potential of how young adult Gen Z women may engage in spaces of online public discourse when the moderation system is rooted in a relational theory of free speech.

In engaging a relational theory of free speech in content moderation systems, moderators are encouraged to view each post in its broader context. Relational theories of free speech emphasize the relationship between the speaker and the intended audience, highlighting how the relationship between both parties shapes the conversation between them. For instance, Jovy Chan's 'free speech as a two-way-right' concept [251] argues that both the speaker and the intended recipient must be considered when evaluating whether the speech in question is 'free' or not. Should one party be coerced or forced into communication with the other, the speech should no longer considered 'free'. Relational theories of speech also emphasize the situation and context surrounding speech. Bonotti and Seglow's relational theory of free speech [252] states that for speech to be considered free it must

ensure that the intended recipient can autonomously speak, the intended recipient has the opportunity to speak, and the speaker is aware that they are accountable and answerable for their words. For example, if two individuals are in a discussion, the speech may cease to become free when one individual no longer feels they can speak without experiencing fear. They no longer have the 'opportunity' to freely and openly express themselves. Furthermore, relational theories of free speech draw attention to how the relationship between a 'speaker' and 'audience' changes over time. Specifically, this relationship is dynamic, with a 'speaker' becoming the 'audience' and vice versa as they communicate. The dynamic nature of this relationship between both parties also subsequently affects the perception of *who* is engaging in 'free speech'. Unlike the culturally pervasive interpretation of free speech that centers on the speaker, relational theories do not center on one specific party. This interpretation of free speech acknowledges that a 'speaker' is not a static role and that a 'speaker' is continuously shifting between being the 'speaker' and being the 'intended audience'. Agora draws on the concepts of relational theories of speech in its design and moderation process.

The Agora community encourages both the speaker (original poster) and the intended audience to be present for discussions surrounding whether a post should be moderated. The relationship and interactions between both are central to the moderation process, as a proper 'evaluation' of the post is dependent on the dialogue between the parties and the context surrounding the post. Agora seeks to create a space for civil discussion between both parties involved in the post. Furthermore, if one party does not reply to use the Agora service, the party who has joined must supply more contextual information in order to ensure a thorough evaluation of the post. If one party explicitly declines to use the service, the post will not be considered by the Agora community. Engaging Chan's 'free speech as a two-way right,' Agora will not force someone to enter a discussion if they would not like to be.

Additionally, by encouraging both the original poster and the intended audience to be

present for the discussion, Agora also draws on the dynamic relationship of a 'speaker' and 'audience' involved in speech. The roles of 'original poster' and 'intended audience' are static within the context of the original post, however, when these individuals are engaged in a dialogue on Agora, they continuously shift between 'speaker' and 'audience'. This process of dialogue between the parties highlights how no one party is central to the relational interpretation of free speech. Speech is not only 'just words' made by the speaker, but also the speaker is not only 'the speaker'. They become 'the audience' at one point during the dialogue and then may shift back to being 'the speaker'.

Voting members of the Agora community consider three questions, largely inspired by concepts from Chan [251] and Bonotti and Seglow [252], when evaluating each post up for moderation. The three points encourage voters to consider the relationship between the original poster and the intended audience and the context in which the post was made. Furthermore, Agora encourages members and posters to communicate with one another when evaluating the post. Original posters and the intended audience may provide additional context to explain their perspective and provide relevant information, enabling voters to have a broader perspective of the context in which the post was made.

6.3.2 Engaging findings from workshop study

In addition to drawing on core concepts of relational theories of free speech, the design of Agora draws on the findings from the workshop study presented in Chapter 5. Participants of the workshop study advocated for more human intervention, negotiation, reflection, and reflexivity during platform moderation processes. They also expressed hesitation at the thought of having to provide platforms with specific details about their offline lives for a more robust and thorough moderation process. Agora draws on the findings from these workshops in parallel with the concepts of relational theories of free speech in its design.

Human Intervention

In this design, human members of the Agora moderation community evaluate posts presented to them by social media platform users. Each post is evaluated by 50 community members, ensuring a thorough review of the post, the context of the post, and the additional context provided by the original poster. The parties involved with the original post may interact with the voting community members to answer questions and provide more information if necessary.

Participants in the workshop study expressed a desire for more human intervention in platform moderation mechanisms. They advocated for more human moderators and multiple rounds of review during the moderation process. Participants believed that by having more human moderators and multiple rounds of review, the context surrounding content would be more thoroughly evaluated, and there would be less room for error. This thought also supports relational theories of free speech that call for a closer examination of context and relationships. In this design fiction, I aim to address this through the community-based moderation approach in which multiple human users evaluate posts and contextual information presented to them.

Negotiation

Posters are encouraged to present the situation and the original post when requesting members of Agora to evaluate whether and how the post should be moderated. Additionally, both the original poster and the intended audience of the post are encouraged to engage in a dialogue with each other and with community members evaluating the post. The emphasis on negotiation and dialogue draws on findings from the workshop study in which participants expressed a desire for mechanisms to negotiate and discuss moderation decisions. Participants called for space on platforms to explain their intentions behind posts or explain why they believe a post should be moderated. Their desire to engage in a dialogue with moderators indicated a broader desire to be more involved in the moderation process.

Drawing on the interactional components between the speaker and the audience that are central to relational theories of speech, Agora creates a space for dialogue between the parties involved in the communication and the community members evaluating the post.

Reflection & Reflexivity

A significant part of the Agora posting process requires that both the original poster and the intended audience reflect on the post. Specifically, the original poster is requested to think through the intention behind the post and their perception of how the intended audience may have interpreted the post. If the post is more indirect, the original poster must speak to their thought process, if any, behind their word choice, use of metaphors, etc. The intended audience is also requested to think through their reaction and articulate how they interpreted the post. For instance, an intended audience may explain why they felt that the post was harmful and or threatening to them, focusing on what words in the original post and what about the context in which it was made contributed to their feeling of potential harm. These points of reflection aim to encourage both the original poster and the intended audience to understand their position within the speech between them.

Similarly, members who are evaluating the post are requested to consider three specific questions in their assessment (See Section 6.2.3). These three points draw on the concepts of relational theories of speech that emphasize understanding the relationship between speaker and audience, and how they interpret the speech between them.

The decision to incorporate points of reflection into the Agora moderation system stems from findings from the workshop study. Participants in the workshop study shared ideas for moderation systems that encouraged users to preemptively reflect on how their posts may impact other users. Although the Agora moderation system works with posts that have already been posted, the system facilitates reflection among both the original poster and the intended audience to enable in-depth dialogue between both parties to better understand how the post should be moderated, if at all.

Desire for Context

In relational theories of free speech, the situated context of the speech shapes the nature of the interactions between the parties involved in the speech. Drawing on this concept, Agora's moderation system requests the parties involved to share as much information as they are comfortable with to contextualize the post for the broader Agora community. The more information voting members of Agora have regarding the post, the more thorough their evaluation of how the post should be moderated. Information regarding when the post was made, whether it was in response to another post, the original poster's relation to the intended audience of the post, how the intended audience interpreted the post targeting them, and the original poster's intention behind the post are all contextually relevant to evaluating if and how the post should be moderated.

Participants of the workshop study presented in Chapter 5 indicated both a desire and hesitation surrounding sharing contextual information with a moderation system. While they agreed that moderation systems should thoroughly investigate the context around a post when evaluating if and how it should be moderated, many expressed privacy concerns. Participants felt that gathering such personal information was out of the scope of a moderation system, and some felt uncomfortable with the thought of moderation systems passing judgment on users and their relationships.

Drawing on this finding, Agora enables the parties involved to choose what information and how much information they share. Since the parties involved may feel uncomfortable revealing information about themselves and their relationships with the community, they may choose what information and how much information they reveal. Additionally, all parties involved must anonymize themselves and share information in a 'generic manner' to not reveal specific details about their lives. One may argue that this may lead posters to selectively disclose information that may bias community members one way or another. While this may be true, the relational theory of free speech underlying this system aims to enable both parties involved to share their perspective with the hope that this leads to a

thoughtful dialogue and, subsequently, a thorough evaluation of the post itself.

6.3.3 Structural Design Inspiration

The term 'Agora' comes from the Greek word for 'assembly'. In ancient Greece, agora referred to a public space where people would gather, discuss, and mingle. In the recent past, the term has also been used to describe social media platforms and their purpose of being a space where people would gather, discuss, and mingle [262, 263].

The inspiration for the design of the Agora moderation system also broadly draws on the 'Forum' style of public discourse in ancient Rome and the democratic voting system. Agora provides a public space for civil dialogue and discussion, similar to the Forum in Rome. The Forum in Rome was not stylized as a space for a public trial but rather a space for community debate in which multiple people could voice their opinions and perspectives [264]. Furthermore, Agora's voting system abides by the general democratic voting system in which the majority opinion is considered the final judgment.

The rules and structure of the Agora moderation system are inspired by the popular subreddit, r/AmITheAsshole [265]. Subreddits on the social media platform Reddit, are often styled as open online forums. In the r/AmITheAsshole subreddit, users make a post explaining a specific conflict and request the larger community to determine who is the 'asshole' in the situation. In their post, users anonymize themselves and others described in the story, revealing only information that is specific and necessary to make a determination. Typically, only one party involved in the conflict presents the story, which often leads to a partial perspective. The structure of the voting guide and rules of Agora are heavily inspired by the ones present in the r/AmITheAsshole subreddit.

This subreddit was chosen for inspiration due to its unique structure and patterns of engagement. Users bring interpersonal issues to the broader community, presenting them in an anonymous story format with the relevant details and then pose the question 'Am I the Asshole?'. Members of the community read through the post and then offer their

perspectives while making a judgment of whether the original poster is the ‘asshole’ in the situation or not. The judgment with the most votes is the one decreed to the original poster.

6.3.4 Limitations of Agora

Agora is just one instantiation of how a relational theory of free speech may be incorporated into the design of a social media platform practice. While the fictional moderation system incorporates several themes central to relational interpretations of free speech, it does have many limitations in its approach. One, Agora solely focuses on addressing directly targeted speech between parties. It does not address speech that is made in a ’1-to-many’ format, which occurs more often than not on social media platforms [266]. danah boyd conceptualizes social media sites as ’networked publics’, explaining that communication often occurs in public or semi-public ways on social media sites, shaping how communication occurs and is broadly perceived by an ’imagined audience’. They further explain that this public/semi-public nature of discourse, compounded with the challenges of ’context-collapse’ and blurred boundaries between the private and public, contribute to the creation of a unique dynamic of participation in online spaces [266]. Concerning hateful, toxic behaviors, the effects of this behavior are amplified due to the ’networked publics’ structure of social media sites. The public nature of toxicity may not only directly affect the individual or group being targeted but also all of those who come across the content. Unfortunately, Agora, at present, does not factor in this prevalent form of participation through its moderation practice. This was due to the limited scope of the design fiction. Agora serves as just one example of how a relational theory of free speech may be incorporated into the underlying theory of a platform moderation practice. Limiting the scope of the types of discourse and dialogue the fictional moderation service addresses served to better illustrate the potential of incorporating the relational interpretation of free speech underlying the moderation service.

Another major limitation of the current design of Agora is that it does not address the

issue of self-censorship. In Chapter 3 of this dissertation, the findings from an interview study with young adult Gen Z women indicate that many of them self-censor and avoid public forms of discussion and discourse to avoid hostility and argumentation online. Many of them choose to do so out of fear for their physical safety, emotional well-being, and or professional safety. Therefore, the public forum nature of Agora may not appeal to this audience. Furthermore, there remains the possibility that members of Agora disseminate violent, hateful, and or false speech. The space open for dialogue and discussion may be misused or weaponized to further target a post's original intended audience. While there are mechanisms in place to thoroughly address any instances of hateful, violent, and or retaliatory behaviors, there is the potential that participants of Agora self-censor and or become argumentative. However, the purpose of Agora is to illustrate the possibility of *what a space can be* when rooted in an alternative interpretation of free speech. One may hope that through repeated exposure to spaces where toxic behaviors that serve to silence open discussion are effectively mitigated, participants will feel less of a need to self-censor.

6.4 Discussion

As aforementioned, Agora is one specific instance of how platform practices may be rooted in relational theories of free speech. Agora draws on and incorporates themes that have been raised and discussed throughout this dissertation, but due to scope, Agora does not aim to address all the potential characteristics of relational theories of free speech. Throughout this section, I will be drawing parallels between themes that have been raised in earlier chapters of this dissertation and themes present in Agora.

6.4.1 Comparison between US-interpretation and relational interpretations of free speech

There are several distinctions between the US-interpretation and relational interpretations of free speech, which are highlighted in their differing approaches to the question of how to moderate speech. I engage existing content moderation practices on platforms such

as Instagram and X as a model of how the culturally pervasive, US-interpretation of free speech is engaged online (See Chapter 4) and compare it to the moderation practices of Agora that have been designed with concepts from relational free speech theories [251, 252, 249].

Focus on Speaker vs. the Relationship

The most significant difference between both theories of free speech is their interpretation of speech between multiple parties. The US-interpretation of free speech is individualized, emphasizing the speaker and their interpretations while ignoring those of the intended audience. In relational theories of free speech, the speaker and intended audience are equally highlighted, with both of their interpretations of speech being accounted for when evaluating the speech. For example, if the speech does not enable the intended audience to reply without fear of potential physical ramifications, then such speech is not considered free [252]. In relational theories of free speech, one's speech is only free if it enables other free speech. This contrasts with the US-interpretation of free speech, which considers speech to be free regardless of its impact or influence on the intended audience. One's relationship to the intended audience is not considered when evaluating whether the speech is free or not. Instead, this interpretation focuses on enabling the speaker to speak freely without consequence. In this interpretation, the 'speaker' is always the 'speaker', and the 'audience' is always the 'audience'. The ability of the 'audience' to become a 'speaker' and engage in their own free speech is not accounted for.

In the context of content moderation, these differing interpretations of free speech affect the underlying approach to moderation. As discussed in Chapter 4, the US-interpretation of free speech underlying existing social media platforms leads platforms to view speech as 'just words', disregarding their broader impact and potential harm on the intended audience. Western Libertarian thought, which largely shapes the US free speech view, supports individualism. As such, this interpretation narrowly focuses on the speaker rather than how

the intended audience is impacted by the speech. Content moderation policies center the speaker, focusing on what they are allowed and not allowed to do or say. For instance, hate speech or disinformation policies describe the types of speech that speakers cannot make, and what may occur to their accounts should they violate the policy [70, 71]. However, there are not many resources or policies in place for targets of hate speech or disinformation who may want speech targeting them to be moderated. For example, if an individual is targeted with gendered disinformation, the only form of recourse they have is to report the speech. Existing policies do not further explain what may happen following the report, or how long it may take to hear back from the report. It is evident that there is a power imbalance between 'speakers' and 'recipients of speech' in these platform policies, further indicating how the policy centers and prioritizes the needs of the 'speaker' and the underlying US-interpretation of free speech.

In contrast, relational theories of free speech center the relationship between both the speaker and their intended audience, and policies with this theory underlying them do the same. It is important to note that relational theories do not highlight the audience's perspective over the speaker's. These theories emphasize that both perspectives need to be accounted for. In Agora, speakers (original posters) and their intended audience are both enabled to voice their perspectives of what they perceive the speech to mean. The intended audience is encouraged to share why they feel that the speech should or should not be moderated, and to share contextual information that supports their perspective. Original posters are similarly encouraged to do the same. The Agora moderation system does not center one party's perspective or needs over the other.

The concept of 'speech as just words' within the US-interpretation of free speech often serves to trivialize a speaker's toxic behavior on online platforms. If speech is just words, then words of harm seemingly do not have any effect on their intended audience and can be disregarded. Policies with this underlying interpretation provide more avenues to protect speakers over their intended audiences, enabling speakers engaging in toxic behavior

to continue without consequence. Under relational theories of free speech, speech is considered a 'relationship', centering the interactional components of speech. Speech is not just words, but, as part of communication, is a mechanism by which a relationship forms between the parties involved. Under such an interpretation, words of harm by one party do affect the other party, and their perspective should be incorporated into evaluating whether that speech was 'free' or not.

Intentionality & Credibility

Another significant component of the US-interpretation of free speech and the evaluation process of whether speech should be considered 'free' is evaluating the speaker's 'intention' and 'credibility'.

As noted in Chapter 4, evaluating speaker intention is fundamental to the moderation of hateful, harmful, and or violent speech on Meta's platforms and X. Meta and X's policies indicate that speech with explicit references to physical harm is considered 'threatening' as it implicates the speaker as the one to want to commit harm. For example, phrases such as 'I will kill you', or 'I am going to hit you with my car' implicate the speaker as the individual with the intention of wanting to commit an act of physical harm. Speech that does not have explicit references to harm is not considered threatening, regardless of whether the person being targeted felt they were in danger. For example, phrases such as 'You better watch your back' or 'I will find you', do not contain references to physical harm and do not implicate the speaker as one to want to commit physical harm. The speaker's intention is unclear in these statements. The speaker may intend to commit an act of physical harm, or may simply intend to warn the user. This notion of 'intention' directly relates to interpretations of what is considered 'credible' [222, 224]. Under this US-interpretation of free speech, a statement may be considered 'credibly threatening' if it contains references to explicit speaker intention. For instance, violent speech moderation policies on Meta and X indicate that a statement such as 'I am going to hit you with my

car' implicates the speaker as one to want to cause harm, and the inclusion of explicit intention makes the statement more 'credibly threatening'. There are other components of the statement that may be considered when evaluating 'credibility', such as whether the statement includes specific details of the acts of harm the speaker intends to commit [71, 70]. However, many of these components are dependent on moderators evaluating the words in the statement rather than the targeted audience's reaction to the statement targeting them. The intended audience's reaction to being targeted by a violent statement is not part of the process of evaluating whether the violent statement is 'credibly threatening' and should be moderated.

While speaker intention and credibility are components of a relational theory of free speech, the process to evaluate whether speech should be considered 'free' does not hinge on the inclusion of explicit speaker intentionality or evidence of credibility. Since relational theories of free speech focus on understanding the interactional components between the speaker and their audience, the perspectives of both parties would be central to the evaluation of whether the speech in question should be considered 'free'. In Agora, an original poster needs to explain their intention behind making the post and what they perceived their audience would feel in reaction to the post. Additionally, the intended audience of the post needs to share their reaction to the post.

One may argue that in this evaluation process, there is an opportunity for both the original poster and the intended audience to omit or obfuscate their true intention or reaction. For instance, an original poster may state that they did not intend for the post to cause harm and that it was meant to be interpreted as a joke. Alternatively, a targeted recipient may argue that the post threatened or harmed them when it did not. While this may be true, Agora intends to provide a space for both the original poster and the intended audience to voice their perspectives on the post and on their interaction and engage in a dialogue about it. Community members can only evaluate the post using the information provided to them. However, if they are provided information from both the original poster and the intended

audience, there is an opportunity to discern a more thorough understanding of the post and how it should or should not be moderated.

Furthermore, the issue of omission and obfuscation exists with the US-interpretation of free speech and within existing platform moderation systems. In Chapter 4, I present how Meta and X's moderation mechanisms go about trying to prove 'intention' and ascertain evidence for 'credibility'. It is difficult to prove 'true intention' as speakers may always plausibly deny that they intended to cause harm through their speech. Speakers may always argue that their speech was 'hyperbolic' or was intended to be humorous. Additionally, it is just as difficult to 'prove' whether a targeted recipient was 'credibly threatened or harmed' by the speech targeting them. However, these existing platform moderation systems only consult one party (the speaker/original poster) when making these decisions, thus always making their judgment based on partial information. Agora may not completely address the issue of omission or obfuscation, however, the fictional moderation service does aim to encourage a more in-depth, contextual evaluation of a post.

Chilling Effects & Slippery Slope

Many advocates of the US-interpretation of free speech justify the need to limit moderation and censorship of offensive, hateful, and harmful speech due to 'chilling effects'. To do this, they engage in the slippery slope argument. As previously discussed in Chapter 4, the chilling effect phenomenon occurs when an individual chooses not to speak out for fear of facing governmental ramifications for their speech. Advocates of the US-interpretation of free speech argue that any mechanism that limits the 'free flow' of speech may lead to chilling effects [232]. Free speech absolutists often engage the 'slippery slope reasoning' to argue that if there are restrictions placed on 'hateful speech' then it may lead to restrictions on other forms of speech [64, 66, 87]. For example, if neo-Nazi speech is censored, it is possible that those censorship rules can then be exploited to censor anti-Nazi speech. In other words, if one form of speech is censored then it may lead to the censorship of many

other forms of speech, which would lead to a chilling effect on speech. While the original definition of 'chilling effects' is focused on governmental interference in speech, the term has been extended to describe forced self-censorship in online contexts [67]. Many Republican leaders in the U.S. have repeatedly claimed that social media moderation policies are biased toward stifling conservative speech and forcing self-censorship among conservatives [30, 29, 267].

Relational theories of speech do speak to the concept of chilling effects, however, they focus on how harmful forms of speech may lead to chilling effects. For example, in their relational theory of free speech, Bonotti and Seglow argue that hate speech should not be considered free speech, since its existence may lead the targeted audience to no longer speak freely out of fear of facing physical ramifications, thus 'chilling' their speech [252]. In other words, the existence of hate speech chills the speech of those who are targeted by it. In her feminist relational theory approach, legal scholar Susan Williams extends this argument, agreeing that hate speech leads to a chilling effect and therefore loses its 'claim to First Amendment value' by US notions of free speech as well. Williams argues that if the purpose of 'free speech' is to discover the truth, as stated in Mill's truth theory [7], then hate speech should not be considered 'free speech' since it prevents those who have been harmed by hate speech from participating in the discovery of the truth [10]. Unlike the US-interpretation of free speech that claims that all speech holds value and is necessary to avoid chilling effects, relational theories of free speech argue that speech that serves to stifle or 'chill' other speech should not hold value and not be considered 'free'.

As part of Agora's voting process, voting members are asked to consider three questions, the first of which is: In this situation, does the intended audience now have the opportunity to reply to this speech without fear of physical, emotional, and or financial retaliation? This question is posed to voters to consider whether the speech in question may have caused a chilling effect on the intended audience. If the audience does not have the opportunity to reply without fear, their speech has been 'chilled' and therefore this speech

is no longer ‘free’.

6.5 Potential Implications of Agora

The potential implementation of Agora as a service that operates in tandem with current-day social media platforms may positively impact engagement in spaces of political discourse. While Agora cannot and will not directly address the issue of hostile argumentation in online political spaces, the system’s approach to mitigating speech that serves to strike fear, discriminate, or silence aims to encourage users to feel comfortable speaking their opinion.

While this dissertation focuses primarily on violent or hateful speech, Agora does not aim to mitigate only these forms of speech. Instances of mis/disinformation may also be addressed by the service. For instance, suppose a post containing vaccine disinformation was brought to the community. The original poster would be required to provide information regarding their intent behind posting and how they perceived the post would be interpreted by the broader audience. Furthermore, the potential effects of disseminating this vaccine disinformation would be considered as part of the moderation process of this post. The potential harms and consequences of this post being circulated would be part of the broader context that would be considered when determining whether or not to moderate this post.

Although I can only speculate as to the effects of Agora on young adult Gen Z women’s engagement in online political discourse spaces, I hope that repeated exposure to a system in which the relational and the contextual are key components of the moderation system may gradually positively affect their public engagement in these spaces.

6.6 Beyond Agora

Design fiction, as a method, serves to highlight the potential of alternative ideologies and realities. Drawing on this method to design Agora, I aimed to draw attention to

how re-conceptualizing moderation practices under an alternative interpretation of free speech may impact practices of public discourse. This then raises the question, why is re-conceptualizing moderation practices under an alternative interpretation of free speech important? Or, more broadly, why might challenging the culturally pervasive US framing of free speech in online contexts be important?

At the time of writing this dissertation, wide-sweeping changes have been made by the U.S. federal government to fundamentally affect how people in the United States express themselves and their ideologies [69]. Those living in the United States are observing this critical point, where the culturally pervasive concept of 'free speech' is being exploited by both the government and social media corporations to silence and marginalize voices of resistance in public discourse. Speech that harms, discriminates, and misinforms has been labeled as 'free', whereas speech that expresses, reclaims, and resists is considered 'not free'. It is at this point where I question whether this culturally pervasive concept of 'free speech' is worth defending in the context of social media.

In order to convince people to trust a system, one must show them that the system can respond in a time of need. However, time and time again, social media users have vocalized their distrust of platform moderation practices and their concern that platforms do little to protect the free speech rights of those being targeted by toxic behaviors online [268, 213]. The recent changes made to limit the practice of moderation by Meta [32] and X [33] do little to address user concerns about being targeted by toxicity. Prior work has already shown the influence of social media practices and policies in shaping public forms of political discourse, specifically in how some of these practices and policies serve to marginalize and exclude folks from these spaces [67, 27, 269]. The work of this dissertation both supports and extends this prior work by highlighting the forms of self-censorship many users exercise when attempting to engage with these public spaces of discourse. Though they are not explicitly excluded from participating, the Gen Z young adult women participants of this research voiced concerns of potentially facing unmitigated toxicity for

sharing their perspective in discourse spaces. I question whether they feel they have the ability to engage in free speech if they are concerned about facing physical and professional ramifications for sharing their opinion publicly.

In Chapter 3 (Section 3.5.5), I interrogate whether a lack of public forms of engagement equates to political apathy. I find that the participants of study in that chapter who did not engage publicly in political discourse were not politically disinterested. Instead, they were deterred from publicly engaging in political discourse due to the prevalence of toxic behaviors and their fear of facing social, physical, and financial repercussions for sharing their perspectives publicly. The work of re-conceptualization through Agora in this chapter is not to claim that public forms of political engagement are the most effective and important, nor that young adult Gen Z women should be publicly engaging in political discourse spaces. As discussed in Chapter 3, not engaging in public forms of political engagement does not equate to political disinterest or apathy. Rather, I propose that young adult Gen Z women should be enabled to engage publicly in political discourse spaces if they would like to. But at present, they are not enabled by social media platforms to do so, compromising their ability to engage in 'free speech' in these spaces. They should be enabled to choose to engage or not engage publicly in these spaces, not be forced to self-censor to protect themselves physically, mentally, and or financially. With the new changes made to platform moderation practices in 2025 [32], it appears that the culturally pervasive interpretation of free speech has been manipulated to serve the conservative agenda of amplifying hostility and polarization. The system will not respond in a time of need, when one's speech is targeted by hateful and toxic behaviors. Instead, it will protect the one who engaged in the hateful, toxic behavior, indicating whose speech platforms consider 'free'.

6.6.1 Engaging Re-conceptualization

The writing of this dissertation served as a mechanism to not only question this culturally pervasive US-interpretation of free speech but also to rethink what 'free speech' really

means in the context of online public discourse. Who, online, is able to truly engage in 'free speech'? How might platforms, through the effects of their policies and practices, limit one's ability to express oneself? This specific moment in time draws attention to how sociopolitical concepts such as 'free speech', which are fundamental to how we engage as a society, are subject to being manipulated by the broader powers that be. For researchers and scholars, this may be a time to reframe the definitions and the values underlying these concepts and, by proxy, these technologies. Many scholars in HCI and CSCW have already done work investigating broad concepts of privacy and community to understand how interactions with technologies may shape and influence public perceptions of these concepts [270]. But how might we resist and counteract the influence of larger corporations in shaping societal perspectives of questions like 'what does it mean to engage in free speech?' or 'what does it mean to express myself?'. Some may argue that developing new designs or creating new platforms and spaces are methods of countering these widespread influences. However, the research of this dissertation points towards a more fundamental change, challenging how we, as a society, interpret 'free speech'. Legal and philosophical scholars have long been questioning whether the culturally pervasive US version of free speech should be applied in the present-day public discourse environment [21, 23, 66, 246, 80]. While these questions are directly related to U.S. law, we as researchers in HCI and social computing can abstract from these discussions and interrogate these questions from an alternative lens. However, the siloes between legal and philosophical scholarship and HCI research limit the potential of cross-disciplinary research and design work that can be done to reframe our interactions with online platforms.

The work of this dissertation identifies an opportunity for cross-disciplinary research that investigates 1) how we may re-conceptualize platform practices using alternative interpretations of concepts such as free speech that challenge the pervasive interpretations, and 2) how these re-conceptualizations may be engaged to resist the dominant interpretations by the government and larger corporations that shape public perceptions. Perhaps through

this process of re-conceptualization, we, as researchers in law, political science, HCI, and STS can create spaces of public discourse that encourage open expression more effectively.

6.6.2 Future Research Directions

In addition to drawing attention to the concept of 'free speech' in social media practices and its potential influence in shaping forms of engagement among young adult Gen Z women in political discourse online, this dissertation broadly serves to initiate a cross-disciplinary examination into how we may challenge and change the underlying interpretation of free speech that dominates and shapes public spaces of online political discourse. Future work in this space may broadly explore 1) alternative interpretations of free speech and how they may be incorporated into the design of social media public discourse practices and 2) alternative re-conceptualization mechanisms to aid existing social media users to question and challenge the dominant interpretation of free speech.

Alternative Interpretations of Free Speech

While this dissertation focused on exploring relational theories of free speech as an alternative to the US-interpretation, there are many other theories of free speech to explore. For example, there are interpretations of free speech that draw on feminist theory [76, 271] and critical race theory [272]. There are interpretations of free speech that are rooted in consequentialism [273], are specific to the context of political speech [274], or are developed due to concerns regarding how the Internet may affect societal understandings of speech [275]. These different interpretations of free speech bring attention to the different dimensions that speech and communication encompass. Furthermore, many scholars in fields such as Communication Theory and Computer-Mediated Communication have theorized alternative interpretations of how communication should be perceived [211, 91]. This scholarship often occurs separately from work done on theorizing 'free speech' despite there being much overlap between the different areas of scholarship. Engaging these different areas of

research in the context of online speech in public political discourse may expand the potential alternative ways that speech and free speech are interpreted on social media platforms.

Furthermore, researchers may also identify specific dimensions or characteristics of these theories of free speech to re-conceptualize. For instance, in this work, I focus on intentionality, credibility, and speaker-centrism as particular dimensions of the US-interpretation of free speech. In Chapter 4, I discuss how these characteristics emerge in moderation practices, and then in this chapter, I again highlight how relational theories of free speech differ in their approach to intentionality, credibility, and speaker-centrism. I draw attention to the differences in both interpretations of free speech through their interpretations of these specific characteristics of a free speech theory. Researchers may similarly begin their process of re-conceptualizing social media practices under alternative interpretations of free speech by first identifying how these alternative interpretations differ on issues of intentionality, credibility, and speaker-centrism. They may then use this as a starting point to rethink moderation practices. For instance, how might a moderation practice differ if the alternative interpretation of free speech did not account for speaker intentionality at all? How might a moderation practice differ if the alternative interpretation of free speech focused on speaker credibility rather than audience credibility?

Alternative Ways to Re-Conceptualize

As aforementioned, through the process of re-conceptualization, researchers are able to highlight the potential of 'what could be'. Not only can researchers engage methods to re-conceptualize among themselves, but they may also use methods to re-conceptualize in order to socialize and promote new ideas. In this chapter, I engage design fiction as an exercise to re-conceptualize moderation practices, however, there is an opportunity to engage other design and research methods to re-conceptualize in future work. Similar methods, such as speculative design [276] or critical design [277], may be effective in promoting the re-conceptualization of social media spaces under alternative theories of free speech. Ad-

ditionally, methods such as action research [278] or critical making [279] in which participants are encouraged to build or design as a way of thinking through critical concepts, may aid researchers. These methods may be used to provoke, reframe, and rethink concepts. There is an opportunity to also engage these methods with scholars in law, policy, communication theory, and philosophy to re-conceptualize social media spaces of discourse from alternative lenses. For instance, asking a free speech scholar to engage speculative design to re-conceptualize how political speech occurs publicly online may reveal certain nuances in how 'free' political speech is perceived compared to hate speech.

6.7 Conclusion

Agora serves as a tool to imagine how content moderation systems may operate under an alternative theory of free speech. How might rooting moderation practices in a relational free speech theory, in which the perspectives of both the speaker and their intended audience are considered when evaluating the speech, re-conceptualize the process of moderating online speech? It is difficult to speculate how young adult Gen Z women may engage in a public space of discourse with such a moderation system without having evaluated this design fiction with them. However, the emphasis on civil discussion and understanding the interactional components of the relationship between speaker and audience may positively affect their public engagement. Participants of the studies presented in Chapters 3 & 5 expressed a deep distrust and skepticism regarding social media as a space for engaging in political discourse; however, much of this distrust was rooted in being continuously exposed to toxic online behavior in political discourse and a sense that their ability to express themselves openly was being restricted. One may speculate that, provided the opportunity to engage in an environment that views speech as relational, their forms of engagement may change.

Agora is not a proposed solution to the issue of disengagement among young adult Gen Z women or even a solution to the limiting nature of free speech absolutism that exists on

social media platforms. Agora is the result of a design fiction thought exercise to explore how relational theories of free speech may present in platform practices.

In the following chapter, I conclude this dissertation by discussing the current state of content moderation practices on social media platforms and the significance of why reframing moderation practices under an alternative theory of free speech may be important in enabling open and public spaces of political discourse.

CHAPTER 7
CONCLUSION

In January 2025, Meta, the parent company of Facebook, Instagram, and WhatsApp, announced that it would cease its fact-checking operations on these platforms. The CEO of Meta, Mark Zuckerberg, cited a desire to increase free speech on the platform as a primary reason to remove its policies on mis/disinformation and generally limit its content moderation practices [32, 280]. Many of these mis/disinformation policies were first implemented following the 2016 U.S. Presidential election when reports indicated that false narratives disseminated on Facebook significantly impacted voting processes and election outcomes [25, 281]. These policies were largely strengthened during the COVID-19 pandemic when harmful mis/disinformation regarding the disease and vaccinations were disseminated across Facebook and Instagram [282]. While many have critiqued Meta's recent changes [283, 284], many have also applauded the company for its 'commitment to the protection of free speech' [280]. Some have argued that the company's decisions stem from past criticism that they are 'anti-Republican' or 'anti-conservative' [267, 285]. While there is a lack of evidence to support these claims, the company's decision to cease fact-checking will hold extensive ramifications for the public sphere of political discourse [29].

Meta's recent moderation changes indicate that the landscape of sociopolitical online discourse is dynamic. One may argue that this is natural. Language is dynamic and constantly shifting, therefore it makes sense that moderation policies also constantly shift. However, the company's recent decision also indicates that discourse that is initially welcomed and considered 'free' may one day no longer be, depending on the whim of the company. Meta's recent changes demonstrate that not only is US free speech absolutism at the forefront of their understanding of what speech is 'free', but also that a few people at the top of a company determine how political discourse occurs for the public.

Platforms and their moderation policies may not only 'shape the shape of discourse', as stated by internet scholar Tarleton Gillespie [16], but also they may shape the shape of what we in U.S. society consider the very concept of 'free speech'. Platform practices and policies shape sociocultural understandings of 'free speech'. One may go as far as to argue that platform policies, through their general inconsistency and unreliability [101, 237], have demonstrated that any online speech can be considered 'free' given the circumstance. Broadly speaking, these platform practices and policies may have inadvertently demonstrated that free speech cannot be considered an 'absolute right' for all of those living in the United States. First Amendment scholar Kyle Langvardt notes, "There is a danger that, after years of acquaintance with content moderation norms on social media, the public will be so inured to them that appeals to free speech norms will not resonate as deeply" [4]. For many, the connotation of the term 'free speech' indicates a right-wing perspective [2, 96], and the right to it may seemingly only apply to users who are vocal in their right-wing rhetoric.

When platforms such as Facebook and Instagram indicate that free speech absolutism, which does not advocate for the free speech of all [66], is central to their platform philosophy, it signals to those who have been marginalized and pushed out of spaces of public discourse that these platforms have never served them. These platforms that have long presented themselves as a space for 'all' have cemented the fact that they do not serve those who are often silenced by the 'survival of the loudest' interpretation of free speech. To borrow from feminist legal scholar Mary Anne Franks, these platforms have embraced the 'Orwellian inversion' in which 'silencing is free speech, and the resistance to silencing is censorship' [66].

Additionally, the fact that the power to make important decisions that determine the free flow of information and public discourse is held by a few indicates that the state of free speech is rather fragile in the U.S. In 2021, following the January 6th attack on Capitol Hill, the tech oligarchy that is Apple, Google, and Amazon disabled access to Parler, a far-right

social media platform. Regardless of one's opinion of the decision, it is important to note that 1) this decision was made quickly, and 2) it was made by essentially 3 people. Many legal and technology scholars noted how the decision of only three CEOs completely upended a public space of discourse, restricting those users' right to freely express themselves in that space [286, 287, 31].

It is at such a point where the question of fundamentally challenging and reframing the interpretation of free speech underlying these platforms is raised. Is it still working? Is it serving the needs of those it promises to serve?

In this dissertation, I explored how the US-interpretation of free speech manifests in online platform practices and policies and how it may shape young adult Gen Z women's perceptions of and engagement in spaces of political discourse. I begin by investigating how young adult Gen Z women engage in political discourse online. I find that young adult Gen Z women are broadly skeptical of using social media as a space to discuss political discourse and often disengage from political discourse and content online. They attribute their disengagement to their fear of facing potential physical, personal, and or professional repercussions for sharing their political opinions online. I then examine underlying interpretations of free speech in social media practices and policies and the role it may play in shaping young adult Gen Z women's forms of engagement online. Through close readings, I find that social media platform policies center the 'speaker' during the moderation process. Explicit speaker intentionality and evidence of 'credibility' are key components to the evaluation process, while the reaction of the individual being targeted by the speech is not considered. The 'speaker-first' approach to platform policies draws on the US-interpretation of free speech that also emphasizes the needs of the speaker over those of the 'hearer' [211, 66].

I then build on the findings from these explorations to understand how engaging an alternative, relational theory of free speech in social media platform practices and policies may affect and change young adult Gen Z women's engagement in online political dis-

course spaces. Relational theories of free speech focus on identifying the contextual and interactional components of communication between the speaker and the hearer. Under this theory, speech is not considered 'free' if it deters one party from speaking or affects their ability to develop their perspective autonomously [252]. I present findings from a workshop study in which Gen Z women participants explore the potential of engaging a relational theory of free speech in Instagram's moderation policies by ideating on alternative forms of moderation. I find that participants advocate for moderation systems rooted in relational theories of free speech to have mechanisms for negotiation, reflection & reflectivity, in addition to more human intervention. Building on the findings from this study and core concepts of relational theories of free speech, I use design fiction methods to explore how engaging an alternative relational theory of free speech in social media practices and policies may be used to challenge the pervasive US notion of free speech. Using my design fiction exercise, Agora, as a case, I compare the US-interpretation of free speech to a relational interpretation of free speech, drawing attention to how online moderation practices would significantly differ. I cannot speculate on how young adult Gen Z women would interact with this fictional moderation system, as it was not evaluated by them. However, the purpose of this design fiction is to highlight the potential of what may come from engaging an alternative theory of free speech in platform practices and policies. Additionally, I use Agora as an example to illustrate the significance of challenging dominant interpretations of sociopolitical concepts such as 'free speech' in our present-day political context.

In this dissertation, my aim is not to suggest that the First Amendment of the U.S. Constitution is obsolete and that the US-interpretation of free speech is incorrect. My background is not in law, and the questions that I raise and interrogate are not concerning U.S law. Rather, through this work of understanding the role that the US-interpretation of free speech plays in social media platform practices and policies, specifically in the context of young adult Gen Z women and their engagement with political discourse, I identify how the US-interpretation of free speech may be limiting speech online rather than enabling it.

This work extends to online platforms, not the U.S. Constitution.

The existing US-interpretation of free speech that is rooted in capitalistic laissez-faire ideals has shaped our path to where those few in power determine how public discourse occurs, what is said in public discourse, and who engages in public discourse. The findings of this research and the existing state of online public discourse indicate that future work should draw on both legal and social computing scholarship to explore alternative interpretations and understandings of free speech to apply to platform practices and policies. Exploring alternatives and examining how they may be employed in online spaces of discourse may serve to address existing issues of silencing and marginalization in these spaces. The outrage and anger at the recent changes announced by Meta, highlight that many do not subscribe to the Libertarian, free speech-absolutist perspective that dominates the Internet. Thus there is optimism that the ”Orwellian Inversion” [66] that has taken over spaces of public discourse may be inverted once again.

Appendices

REFERENCES

[1] U. Constitution, *First amendment*, Retrieved from https://www.archives.gov/founding-docs/bill-of-rights-transcript, Accessed: 2025-03-06, 1787.

[2] A. Gibson, "Free Speech and Safe Spaces: How Moderation Policies Shape Online Discussion Spaces," *Social Media + Society*, vol. 5, no. 1, p. 2 056 305 119 832 588, Jan. 2019, Publisher: SAGE Publications Ltd.

[3] K. Paul, "Zuckerberg defends Facebook as bastion of 'free expression' in speech," *The Guardian*, Oct. 2019.

[4] K. Langvardt, "Regulating Online Content Moderation," *Georgetown Law Journal*, 2017.

[5] *Freedom of Speech, Not Reach: An update on our enforcement philosophy*, Apr. 2023.

[6] E. T. Kasper and T. A. Kozma, *The Supreme Court and the Philosopher: How John Stuart Mill Shaped US Free Speech Protections*. Cornell University Press, Apr. 2024, Google-Books-ID: G6_oEAAAQBAJ, ISBN: 978-1-5017-7453-9.

[7] J. S. Mill, *On Liberty*. Longmans, Green, Reader, and Dyer, 1869, Google-Books-ID: RbkAAAAAYAAJ, ISBN: 978-0-14-188320-5.

[8] D. Schultz, *Marketplace of Ideas*, 2009.

[9] D. C. Nunziato, "The Marketplace of Ideas Online," *Notre Dame Law Review*, vol. 94, no. 4, 2019.

[10] S. H. Williams, "Feminist Jurisprudence and Free Speech Theory," *Tulane Law Review*, vol. 68, no. 6, 1994.

[11] H. L. Review, "First Amendment — Freedom of Speech — Trademarks — Matal v. Tam," *Harvard Law Review*, vol. 131, no. 1, Nov. 2017, Section: Leading Case.

[12] M. A. Franks, "Beyond the Public Square: Imagining Digital Democracy," *The Yale Law Journal Forum*, vol. 131, Nov. 2021.

[13] Z. Papacharissi, "The virtual sphere: The internet as a public sphere," *New Media & Society*, vol. 4, no. 1, pp. 9–27, Feb. 2002, Publisher: SAGE Publications.

[14] D. Yeung. "The 'Digital Town Square' Problem." (Jan. 2023).

[15] Mark Zuckerberg, *A Privacy-Focused Vision for Social Networking*, Section: Technology, Mar. 2019.

[16] T. Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press, Jun. 2018, Google-Books-ID: cOJgDwAAQBAJ, ISBN: 978-0-300-23502-9.

[17] H.-C. H. Chang, A. Richardson, and E. Ferrara, "#JusticeforGeorgeFloyd: How Instagram facilitated the 2020 Black Lives Matter protests," *PLOS ONE*, vol. 17, no. 12, e0277864, Dec. 2022.

[18] T. H. Chu and T. E. D. Yeo, "Rethinking Social Media and Political Engagement: An Examination of the Disconnective Practices of Politically Active Youths in Hong Kong," in *Proceedings of the 9th International Conference on Social Media and Society*, ser. SMSociety '18, New York, NY, USA: Association for Computing Machinery, Jul. 2018, pp. 41–50, ISBN: 978-1-4503-6334-1.

[19] M. Conroy, J. T. Feezell, and M. Guerrero, "Facebook and political engagement: A study of online political group membership and offline political engagement," *Computers in Human Behavior*, vol. 28, no. 5, pp. 1535–1546, Sep. 2012.

[20] D. Simonovic, "Report of the Special Rapporteur on Violence against Women, Its Causes and Consequences on online violence against women and girls from a human rights perspective," United Nations, Geneva, Tech. Rep., Jun. 2018, Publisher: UN,

[21] T. Wu, "Is the First Amendment Obsolete?" *Michigan Law Review*, vol. 117, 2018.

[22] T. M. Massaro and H. Norton, "Free Speech and Democracy: A Primer for Twenty-First Century Reformers," vol. 54, 2021.

[23] J. Waldron, *The Harm in Hate Speech*. Harvard University Press, 2012.

[24] E. Bazelon, "The Problem of Free Speech in an Age of Disinformation," *The New York Times*, Oct. 2020.

[25] H. Allcott and M. Gentzkow, "Social Media and Fake News in the 2016 Election," *The Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–235, 2017, Publisher: American Economic Association.

[26] D. Thakur and D. L. Hankerson, "Facts and their Discontents: A Research Agenda for Online Disinformation, Race, and Gender," Center for Democracy & Technology, Tech. Rep., Feb. 2021.

[27] D. Thakur and D. L. Hankerson, "An Unrepresentative Democracy: How Disinformation and Online Abuse Hinder Women of Color Political Candidates in the United States," Center for Democracy & Technology, Tech. Rep., Oct. 2022.

[28] V. L. Killion, "The First Amendment: Categories of Speech," Congressional Research Service, Tech. Rep., Jan. 2019.

[29] P. M. Barrett and J. G. Sims, "False Accusation: The Unfounded Claim That Social Media Companies Censor Conservatives," NYU Stern Center for Business & Human Rights, Tech. Rep., Feb. 2021.

[30] P. M. Barrett, *More Evidence that Conservatives Are Not Unfairly Censored on Social Media*, Dec. 2024.

[31] R. Wicker, *Wicker: Big Tech Has Crossed a Red Line*, Section: Weekly Reports, Jan. 2021.

[32] Joel Kaplan, *More Speech and Fewer Mistakes*, Jan. 2025.

[33] S. Bhaimiya, *Twitter will lose over 32 million users by 2024 as Elon Musk's takeover causes more hate speech and technical problems, report says*.

[34] C. Stokel-Walker, *Twitter may have lost more than a million users since Elon Musk took over*, Nov. 2022.

[35] T. Bajpai, D. Asher, A. Goswami, and E. Chandrasekharan, "Harmonizing the Cacophony with MIC: An Affordance-aware Framework for Platform Moderation," in *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, Nov. 2022, pp. 1–22.

[36] R. Gorwa, R. Binns, and C. Katzenbach, "Algorithmic content moderation: Technical and political challenges in the automation of platform governance," *Big Data & Society*, vol. 7, no. 1, p. 2 053 951 719 897 945, Jan. 2020, Publisher: SAGE Publications Ltd.

[37] G. Koushik, K. Rajeswari, and S. K. Muthusamy, "Automated Hate Speech Detection on Twitter," in *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, Sep. 2019, pp. 1–4.

[38] D. Hettiachchi and J. Goncalves, "Towards Effective Crowd-Powered Online Content Moderation," in *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*, ser. OZCHI'19, New York, NY, USA: Association for Computing Machinery, Jan. 2020, pp. 342–346, ISBN: 978-1-4503-7696-9.

[39] J. A. Jiang, P. Nie, J. R. Brubaker, and C. Fiesler, "A Trade-off-centered Framework of Content Moderation," in *ACM Transactions on Computer-Human Interaction*, vol. 30, Feb. 2023, pp. 1–34.

[40] M. Chan, H.-T. Chen, and F. L. F. Lee, "Examining the roles of mobile and social media in political participation: A cross-national analysis of three Asian societies using a communication mediation approach," *New Media & Society*, vol. 19, no. 12, pp. 2003–2021, Dec. 2017, Publisher: SAGE Publications.

[41] M. J. Glennon, *Free Speech and Turbulent Freedom: The Dangerous Allure of Censorship in the Digital Era*. Oxford, New York: Oxford University Press, Jan. 2024, ISBN: 978-0-19-763676-3.

[42] Danielle Keats Citron and Jonathon Penney, "Empowering Speech by Moderating It — American Academy of Arts and Sciences," *Daedalus*, Aug. 2024.

[43] A. Rauchfleisch and A. Jungherr, "Blame and obligation: The importance of libertarianism and political orientation in the public assessment of disinformation in the United States," *Policy & Internet*, vol. n/a, no. n/a, May 2024.

[44] B. Pfaffenberger, ""If I Want It, It's OK": Usenet and the (Outer) Limits of Free Speech," *The Information Society*, vol. 12, no. 4, pp. 365–386, Nov. 1996, Publisher: Routledge _eprint: https://doi.org/10.1080/019722496129350.

[45] R. Jiménez Durán, *The Economics of Content Moderation: Theory and Experimental Evidence from Hate Speech on Twitter*, SSRN Scholarly Paper, Rochester, NY, Nov. 2021.

[46] S. C. Herring, "The Rhetorical Dynamics of Gender Harassment On-Line," *The Information Society*, vol. 15, no. 3, pp. 151–167, Aug. 1999, Publisher: Routledge _eprint: https://doi.org/10.1080/019722499128466.

[47] M. Dimock, *Defining generations: Where Millennials end and Generation Z begins*.

[48] L. L. Rice and K. W. Moffett, *The Political Voices of Generation Z*. Routledge, Sep. 2021, Google-Books-ID: yVdBEAAAQBAJ, ISBN: 978-1-00-045034-7.

[49] J. Kahne, N. Lee, and J. Feezell, "The Civic and Political Significance of Online Participatory Cultures among Youth Transitioning to Adulthood," *Journal of Information Technology & Politics*, vol. 10, Jan. 2012.

[50] A. Choi, C. D'Ignazio, B. Foucault Welles, and A. G. Parker, "Social Media as a Critical Pedagogical Tool: Examining the Relationship between Youths' Online Sociopolitical Engagements and Their Critical Consciousness," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI '23,

New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–25, ISBN: 978-1-4503-9421-5.

[51] Soumitra Dutta and Matthew Fraser, "Barack Obama and the Facebook Election," *US News & World Report*, Nov. 2018.

[52] K. Hamza, "Social media as a tool for social movements in Arab spring countries," in *Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance*, Guimaraes Portugal: ACM, Oct. 2014, pp. 71–74, ISBN: 978-1-60558-611-3.

[53] K. Aal, N. Boulus-Rødje, and M. Rohde, "Media, Technologies, Cooperation – Rethinking Publics and Publicness in the MENA Region," *Computer Supported Cooperative Work (CSCW)*, vol. 31, no. 2, pp. 149–157, Jun. 2022.

[54] Z. Sanderson, M. A. Brown, R. Bonneau, J. Nagler, and J. A. Tucker, "Twitter flagged Donald Trump's tweets with election misinformation: They continued to spread both on and off the platform," *Harvard Kennedy School Misinformation Review*, Aug. 2021.

[55] Benjy Sarlin, "'Fake news' went viral in 2016. This professor studied who clicked.," *NBC News*, Jan. 2018.

[56] S. C. Herring, "Gender and democracy in computer-mediated communication," in *Computerization and controversy (2nd ed.)*, 1996.

[57] L. Bode, "Closing the gap: Gender parity in political engagement on social media," *Information, Communication & Society*, vol. 20, no. 4, pp. 587–603, Apr. 2017.

[58] C. Peacock and E. Van Duyn, "Monitoring and correcting: Why women read and men comment online," *Information, Communication & Society*, vol. 26, no. 6, pp. 1106–1121, Apr. 2023.

[59] E. Van Duyn, C. Peacock, and N. J. Stroud, "The Gender Gap in Online News Comment Sections," *Social Science Computer Review*, vol. 39, no. 2, pp. 181–196, Apr. 2021, Publisher: SAGE Publications Inc.

[60] S. Schoenebeck *et al.*, "Online Harassment in Majority Contexts: Examining Harms and Remedies across Countries," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI '23, New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–16, ISBN: 978-1-4503-9421-5.

[61] B. Poland, *Haters: Harassment, Abuse, and Violence Online*. University of Nebraska Press, 2016, ISBN: 978-1-61234-766-0.

[62] J. Im *et al.*, "Women's Perspectives on Harm and Justice after Online Harassment," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, 355:1–355:23, Nov. 2022.

[63] E. A. Jane, *Misogyny Online: A Short (and Brutish) History*. SAGE, Oct. 2016, Google-Books-ID: PxwZDQAAQBAJ, ISBN: 978-1-4739-2715-5.

[64] D. Cole, *Defending Speech We Hate — ACLU*, Feb. 2024.

[65] L. Rowland. "We All Need to Defend Speech We Hate." (Apr. 2017), (visited on 03/22/2023).

[66] M. A. Franks, "Feminist Frictions: Speaking of Women: Feminism and Free Speech," *Signs: Journal of Women in Culture and Society*, 2022.

[67] S. Sobieraj, *Credible Threat: Attacks Against Women Online and the Future of Democracy*. Oxford University Press, Sep. 2020, ISBN: 978-0-19-008930-6.

[68] K. Chadha, L. Steiner, J. Vitak, and Z. Ashktorab, "Women's Responses to Online Harassment," *International Journal of Communication*, vol. 14, no. 0, p. 19, Jan. 2020, Number: 0.

[69] E. Bumiller, "Fearing Retribution, Trump Critics Muzzle Themselves," *The New York Times*, Mar. 2025.

[70] X. H. Center, *Violent Content Policy*, May 2024.

[71] *Violence and Incitement — Transparency Center*.

[72] M. University of Wisconsin, *What is the difference between protected and unprotected speech?*

[73] *What Does Free Speech Mean? — United States Courts*.

[74] A. Satariano and C. F. Schuetze, "Where Online Hate Speech Can Bring the Police to Your Door," *The New York Times*, Sep. 2022.

[75] *United States v. Schwimmer, 279 U.S. 644*, 1929.

[76] S. Williams, *Truth, Autonomy, and Speech: Feminist Theory and the First Amendment*. NYU Press, Jun. 2004, ISBN: 978-0-8147-9511-8.

[77] L. Lessig, *Code and other laws of cyberspace*, Nachdr. New York: Basic Books, 2000, ISBN: 978-0-465-03912-8.

[78] J. L. Short, R. Schiller, S. S. Silbey, N. Jones, B. Hammatian, and L. Bowman-Carpio, "The Dog that Didn't Bark: Looking for Techno-Libertarian Ideology in a Decade of Public Discourse About Big Tech Regulation," *Ohio State Technology Law Journal*, vol. 19, no. 1, 2022.

[79] Meta, *Mark Zuckerberg Stands for Voice and Free Expression*, Oct. 2019.

[80] K. Lasson, "Racial Defamation as Free Speech: Abusing the First Amendment," *Columbia Human Rights Law Review*, vol. 17, no. 1, pp. 11–56, 1985.

[81] C. E. Baker, *Human liberty and freedom of speech*. New York: Oxford University Press, 1989, ISBN: 978-0-19-505777-5.

[82] J. Gordon, "John Stuart Mill and the "Marketplace of Ideas"," *Social Theory and Practice*, vol. 23, no. 2, pp. 235–249, 1997, Publisher: Florida State University Department of Philosophy.

[83] D. Brink, "Mill's Moral and Political Philosophy," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta and U. Nodelman, Eds., Fall 2022, Metaphysics Research Lab, Stanford University, 2022.

[84] L. Williams, *John Stuart Mill*, 2009.

[85] *Snyder v. Phelps 562 U.S. 443*, 2011.

[86] A. M. Lofaso, "Baker's Autonomy Theory of Free Speech," *West Virginia Law Review*, vol. 115, 2012.

[87] E. Baker, *Hate Speech*, SSRN Scholarly Paper, Rochester, NY, Mar. 2008.

[88] *West Virginia State Board of Education v. Barnette 319 U.S 624*, Jun. 1943.

[89] *Stanley v. Georgia, 394 U.S. 557*, Apr. 1969.

[90] *303 Creative LLC v. Elenis, 600 U.S 570*, Jun. 2023.

[91] S. Herring, "Freedom of speech or freedom of harassment?" *The College*, vol. 1, pp. 8–9, Jan. 1995.

[92] J. Reagle, ""Free as in sexist?" Free culture and the gender gap," *First Monday*, 2013.

[93] J. Habermas, *The Structural Transformation of the Public Sphere: An Inquiry Into a Category of Bourgeois Society*. Polity Press, 1989, Google-Books-ID: 3iIVnwEACAAJ, ISBN: 978-0-7456-0274-5.

[94] N. Fraser, "Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy," *Social Text*, no. 25/26, p. 56, 1990.

[95] B. W. Hawkins and O. Haimson, "Building an online community of care: Tumblr use by transgender individuals," in *Proceedings of the 4th Conference on Gender & IT*, ser. GenderIT '18, New York, NY, USA: Association for Computing Machinery, May 2018, pp. 75–77, ISBN: 978-1-4503-5346-5.

[96] A. Marwick and R. Lewis, "Media Manipulation and Disinformation Online," Data & Society, Tech. Rep., 2017.

[97] E. Volokh, *47 U.S.C. § 230 and the Publisher/Distributor/Platform Distinction*, Section: Civil Liberties, May 2020.

[98] Valerie C. Brannon and Eric N. Holmes, "Section 230: An Overview," Congressional Research Service, legislation, Jan. 2024.

[99] E. F. Foundation, *Section 230 and Other Content Moderation Laws Across The Globe*.

[100] *Committee to Hold Hearing with Big Tech CEOs on Section 230*, Section: Press Releases, Oct. 2020.

[101] S. Roberts, "Digital Refuse: Canadian Garbage, Commercial Content Moderation and the Global Circulation of Social Media's Waste," *Wi: Journal of Mobile Media*, Jan. 2016.

[102] Meta, *Facebook Community Standards*.

[103] T. T. Rules, *The Twitter rules: Safety, privacy, authenticity, and more*.

[104] Reddit, *Content Policy*.

[105] S. Singh, *Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content*, Jul. 2019.

[106] *Moderation actions - Mastodon documentation*.

[107] A. George, *Discord now has an automated content moderation tool*, Jun. 2022.

[108] K. Paul and S. Dang, "Exclusive: Twitter leans on automation to moderate content as harmful speech surges," *Reuters*, Dec. 2022.

[109] A. O. Hirschman, *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Cambridge, MA: Harvard University Press, Feb. 1972, ISBN: 978-0-674-27660-4.

[110] A. S. Bruckman, *Should You Believe Wikipedia?: Online Communities and the Construction of Knowledge*. Cambridge: Cambridge University Press, 2022, ISBN: 978-1-108-49032-0.

[111] C. Newton and N. Tiku, *Twitter CEO: 'We suck at dealing with abuse'*, Feb. 2015.

[112] B. Fung, *Twitter prepares to roll out new paid subscription service that includes blue checkmark — CNN Business*, Nov. 2022.

[113] A. Hutchinson, *Twitter Outlines New Platform Rules Which Emphasize Reduced Reach, as Opposed to Suspensions*, Jan. 2023.

[114] Y. Liu, P. Yildirim, and Z. J. Zhang, *Implications of Revenue Models and Technology for Content Moderation Strategies*, SSRN Scholarly Paper, Rochester, NY, Nov. 2021.

[115] L. Madio and M. Quinn, *Content moderation and advertising in social media platforms*, SSRN Scholarly Paper, Rochester, NY, Mar. 2023.

[116] S. Illing, *The First Amendment has a Facebook problem*, Apr. 2021.

[117] Z. Papacharissi, "Democracy online: Civility, politeness, and the democratic potential of online political discussion groups," *New Media & Society*, vol. 6, no. 2, pp. 259–283, Apr. 2004, Publisher: SAGE Publications.

[118] E. Anduiza, A. Gallego, and M. Cantijoch, "Online Political Participation in Spain: The Impact of Traditional and Internet Resources," *Journal of Information Technology & Politics*, vol. 7, no. 4, pp. 356–368, Oct. 2010, Publisher: Routledge _eprint: https://doi.org/10.1080/19331681003791891.

[119] E. Bakshy, S. Messing, and L. A. Adamic, "Exposure to ideologically diverse news and opinion on Facebook," *Science*, vol. 348, no. 6239, pp. 1130–1132, Jun. 2015.

[120] S. Boulianne, "Does Internet Use Affect Engagement? A Meta-Analysis of Research," *Political Communication*, vol. 26, no. 2, pp. 193–211, May 2009, Publisher: Routledge _eprint: https://doi.org/10.1080/10584600902854363.

[121] D. V. Shah, N. Kwak, and R. L. Holbert, ""Connecting" and "Disconnecting" With Civic Life: Patterns of Internet Use and the Production of Social Capital," *Political Communication*, vol. 18, no. 2, pp. 141–162, Apr. 2001.

[122] Z. Papacharissi and M. De Fatima Oliveira, "Affective News and Networked Publics: The Rhythms of News Storytelling on #Egypt," *Journal of Communication*, vol. 62, no. 2, pp. 266–282, Apr. 2012.

[123] K. Starbird and L. Palen, "(How) will the revolution be retweeted? information diffusion and the 2011 Egyptian uprising," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, ser. CSCW '12, New York, NY, USA: Association for Computing Machinery, Feb. 2012, pp. 7–16, ISBN: 978-1-4503-1086-4.

[124] C. Ruess, C. P. Hoffmann, S. Boulianne, and K. Heger, "Online political participation: The evolution of a concept," *Information, Communication & Society*, vol. 26, no. 8, pp. 1495–1512, Jun. 2023.

[125] Y. Theocharis, S. Boulianne, K. Koc-Michalska, and B. Bimber, "Platform affordances and political participation: How social media reshape political engagement," *West European Politics*, vol. 46, no. 4, pp. 788–811, Jun. 2023, Publisher: Routledge _eprint: https://doi.org/10.1080/01402382.2022.2087410.

[126] U. of Sussex, *Is digital activism effective? — University of Sussex*, Section: Blog, Jun. 2022.

[127] M. Nelimarkka, J. P. Rancy, J. Grygiel, and B. Semaan, "(Re)Design to Mitigate Political Polarization: Reflecting Habermas' ideal communication space in the United States of America and Finland," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, 141:1–141:25, Nov. 2019.

[128] K. Grön and M. Nelimarkka, "Party Politics, Values and the Design of Social Media Services: Implications of political elites' values and ideologies to mitigating of political polarisation through design," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, 104:1–104:29, Oct. 2020.

[129] N. Schneider, P. De Filippi, S. Frey, J. Z. Tan, and A. X. Zhang, "Modular Politics: Toward a Governance Layer for Online Communities," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, 16:1–16:26, Apr. 2021.

[130] A. Baughan *et al.*, "Someone Is Wrong on the Internet: Having Hard Conversations in Online Spaces," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, 156:1–156:22, Apr. 2021.

[131] C. Grevet, L. G. Terveen, and E. Gilbert, "Managing political differences in social media," in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, ser. CSCW '14, New York, NY, USA: Association for Computing Machinery, Feb. 2014, pp. 1400–1408, ISBN: 978-1-4503-2540-0.

[132] I. Saric, *Jacinda Ardern's resignation and the toll of burnout and abuse for women leaders*, Jan. 2023.

[133] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis, "Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship," in *Proceedings of the 2018 World Wide Web Conference*, ser. WWW '18, Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, Apr. 2018, pp. 913–922, ISBN: 978-1-4503-5639-8.

[134] E. H. R. Rho, G. Mark, and M. Mazmanian, "Fostering Civil Discourse Online: Linguistic Behavior in Comments of #MeToo Articles across Political Perspectives," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, 147:1–147:28, Nov. 2018.

[135] A. Aldayel and W. Magdy, "Your Stance is Exposed! Analysing Possible Factors for Stance Detection on Social Media," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, 205:1–205:20, Nov. 2019.

[136] J. Mahoney, T. Feltwell, O. Ajuruchi, and S. Lawson, "Constructing the Visual Online Political Self: An Analysis of Instagram Use by the Scottish Electorate," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI '16, New York, NY, USA: Association for Computing Machinery, May 2016, pp. 3339–3351, ISBN: 978-1-4503-3362-7.

[137] A. Velasquez, "Social media and online political discussion: The effect of cues and informational cascades on participation in online political communities," *New Media & Society*, vol. 14, no. 8, pp. 1286–1303, Dec. 2012, Publisher: SAGE Publications.

[138] R. K. Garrett, "Echo chambers online?: Politically motivated selective exposure among Internet news users," *Journal of Computer-Mediated Communication*, vol. 14, no. 2, pp. 265–285, Jan. 2009.

[139] R. Hada *et al.*, "Beyond Digital "Echo Chambers": The Role of Viewpoint Diversity in Political Discussion," in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, ser. WSDM '23, New York, NY, USA: Association for Computing Machinery, Feb. 2023, pp. 33–41, ISBN: 978-1-4503-9407-9.

[140] I. Pentina and M. Tarafdar, "From "information" to "knowing": Exploring the role of social media in contemporary news consumption," *Computers in Human Behavior*, vol. 35, pp. 211–223, Jun. 2014.

[141] K. Conger and L. Hirsch, "Elon Musk Completes $44 Billion Deal to Own Twitter," *The New York Times*, Oct. 2022.

[142] K. Rosenblatt, "Gen Z is over Facebook, Pew Research Center finds," *NBC News*, Aug. 2022.

[143] C. Holtermann, "They're Over Being Real," *The New York Times*, Apr. 2023.

[144] A. P. Kambhampaty, "Sidechat Wants to Be College Students' Main Chat," *The New York Times*, May 2022.

[145] R. J. Bonnie *et al.*, "Introduction," in *Investing in the Health and Well-Being of Young Adults*, National Academies Press (US), Jan. 2015.

[146] A. Blandford, D. Taylor, and M. Smit, "Examining the role of information in the civic engagement of youth," in *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, ser. ASIST '15, USA: American Society for Information Science, Nov. 2015, pp. 1–9, ISBN: 978-0-87715-547-8.

[147] H. Hwang and K.-O. Kim, "Social media as a tool for social movements: The effect of social media use and social capital on intention to participate in social movements," *International Journal of Consumer Studies*, vol. 39, no. 5, pp. 478–488, 2015, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ijcs.12221.

[148] A. Oden and L. Porter, "The Kids Are Online: Teen Social Media Use, Civic Engagement, and Affective Polarization," *Social Media + Society*, vol. 9, no. 3, p. 20 563 051 231 186 364, Jul. 2023, Publisher: SAGE Publications Ltd.

[149] Y. Wang and G. Mark, "Engaging with Political and Social Issues on Facebook in College Life," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, Portland Oregon USA: ACM, Feb. 2017, pp. 433–445, ISBN: 978-1-4503-4335-0.

[150] M. Ekström, "Young people's everyday political talk: A social achievement of democratic engagement," *Journal of Youth Studies*, vol. 19, no. 1, pp. 1–19, Jan. 2016.

[151] E. Middaugh, B. Bowyer, and J. Kahne, "U Suk! Participatory Media and Youth Experiences With Political Discourse," *Youth & Society*, vol. 49, no. 7, pp. 902–922, Oct. 2017, Publisher: SAGE Publications Inc.

[152] K. Thorson, "Facing an uncertain reception: Young citizens and political interaction on Facebook," *Information, Communication & Society*, vol. 17, no. 2, pp. 203–216, Feb. 2014.

[153] E. C. Weinstein, M. Rundle, and C. James, "A Hush Falls Over the Crowd: Diminished Online Civic Expression Among Young Civic Actors," *International Journal of Communication*, vol. 9, no. 0, p. 23, Jan. 2015, Number: 0.

[154] G. Mascheroni and M. F. Murru, ""I Can Share Politics But I Don't Discuss It": Everyday Practices of Political Talk on Facebook," *Social Media + Society*, vol. 3, no. 4, p. 2 056 305 117 747 849, Oct. 2017, Publisher: SAGE Publications Ltd.

[155] J. Penney, "Young People as Political Influencers on Social Media: Skepticism and Network Thinking," in *Proceedings of the 9th International Conference on Social Media and Society*, Copenhagen Denmark: ACM, Jul. 2018, pp. 355–359, ISBN: 978-1-4503-6334-1.

[156] Scott Keeter, Juliana Horowitz, and Alec Tyson, *Young Voters in the 2008 Election*, Nov. 2008.

[157] Y. News, *Number of active users at Facebook over the years*, May 2013.

[158] Scott Nover, "Facebook passed 3 billion users for the first time," *Quartz*, Jul. 2023.

[159] N. Confessore, "Cambridge Analytica and Facebook: The Scandal and the Fallout So Far," *The New York Times*, Apr. 2018.

[160] C. Faith, "Social Media As A Political Platform In Nigeria: A Focus On Electorates In South-Eastern Nigeria," *Eastern Nigeria*, 2016.

[161] P. Mozur, "A Genocide Incited on Facebook, With Posts From Myanmar's Military," *The New York Times*, Oct. 2018.

[162] Barbara Ortutay, "Amnesty report finds Facebook amplified hate ahead of Rohingya massacre in Myanmar," *PBS News*, Sep. 2022, Section: World.

[163] B. Auxier, *Social media continue to be important political outlets for Black Americans*, Dec. 2020.

[164] David Klepper and Amanda Seitz, "Facebook froze as anti-vaccine comments swarmed users," *AP News*, Oct. 2021.

[165] Amanda Seitz and Hannah Fingerhut, "Americans agree misinformation is a problem, poll shows," *AP News*, Oct. 2021.

[166] J. Vitak, K. Chadha, L. Steiner, and Z. Ashktorab, "Identifying Women's Experiences With and Strategies for Mitigating Negative Effects of Online Harassment," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative*

Work and Social Computing, ser. CSCW '17, New York, NY, USA: Association for Computing Machinery, Feb. 2017, pp. 1231–1245, ISBN: 978-1-4503-4335-0.

[167] A. Kaskazi and V. Kitzie, "Engagement at the margins: Investigating how marginalized teens use digital media for political participation," *New Media & Society*, vol. 25, no. 1, pp. 72–94, Jan. 2023, Publisher: SAGE Publications.

[168] R. Das, "Speaking About Birth: Visible and Silenced Narratives in Online Discussions of Childbirth," *Social Media + Society*, vol. 3, no. 4, p. 2 056 305 117 735 753, Oct. 2017, Publisher: SAGE Publications Ltd.

[169] N. Andalibi, O. L. Haimson, M. De Choudhury, and A. Forte, "Understanding Social Media Disclosures of Sexual Abuse Through the Lenses of Support Seeking and Anonymity," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI '16, New York, NY, USA: Association for Computing Machinery, May 2016, pp. 3906–3918, ISBN: 978-1-4503-3362-7.

[170] A. Dashti, H. Al-Abdullah, and H. Johar, "Social Media and the Spiral of Silence: The Case of Kuwaiti Female Students Political Discourse on Twitter," *Journal of International Women's Studies*, vol. 16, no. 3, pp. 42–53, Jul. 2015.

[171] D. K. Brown, Y. M. M. Ng, M. J. Riedl, and I. Lacasa-Mas, "Reddit's Veil of Anonymity: Predictors of engagement and participation in media environments with hostile reputations," *Social Media + Society*, vol. 4, no. 4, Oct. 2018, Publisher: SAGE Publications Ltd.

[172] K. Schulenberg, G. Freeman, L. Li, and C. Barwulor, ""'Creepy Towards My Avatar Body, Creepy Towards My Body": How Women Experience and Manage Harassment Risks in Social Virtual Reality," *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. CSCW2, 236:1–236:29, Oct. 2023.

[173] A. X. Zhang, G. Hugh, and M. S. Bernstein, "PolicyKit: Building Governance in Online Communities," in *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '20, New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 365–378, ISBN: 978-1-4503-7514-6.

[174] S. Sultana *et al.*, "'Unmochon': A Tool to Combat Online Sexual Harassment over Facebook Messenger," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI '21, New York, NY, USA: Association for Computing Machinery, May 2021, pp. 1–18, ISBN: 978-1-4503-8096-6.

[175] J. Sasse and J. Grossklags, "Breaking the Silence: Investigating Which Types of Moderation Reduce Negative Effects of Sexist Social Media Content," *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. CSCW2, 327:1–327:26, Oct. 2023.

[176] C. R. Carlson and H. Witt, "Online harassment of U.S. women journalists and its impact on press freedom," *First Monday*, Oct. 2020.

[177] D. K. Citron, *Hate Crimes in Cyberspace*. Harvard University Press, Sep. 2014, Google-Books-ID: wmiABAAAQBAJ, ISBN: 978-0-674-36829-3.

[178] E. H. R. Rho and M. Mazmanian, "Political Hashtags & the Lost Art of Democratic Discourse," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20, New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 1–13, ISBN: 978-1-4503-6708-0.

[179] S. Yardi and D. Boyd, "Dynamic Debates: An Analysis of Group Polarization Over Time on Twitter," *Bulletin of Science, Technology & Society*, vol. 30, no. 5, pp. 316–327, Oct. 2010, Publisher: SAGE Publications Inc.

[180] F. Irannejad Bisafar, B. Foucault Welles, C. D'Ignazio, and A. G. Parker, "Supporting Youth Activists? Strategic Use of Social Media: A Qualitative Investigation of Design Opportunities," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, 109:1–109:25, Oct. 2020.

[181] T. P. Bakker and C. H. de Vreese, "Good News for the Future? Young People, Internet Use, and Political Participation," *Communication Research*, vol. 38, no. 4, pp. 451–470, Aug. 2011, Publisher: SAGE Publications Inc.

[182] Mark Hall, *Facebook*, Jul. 2024.

[183] Mark Hall and William L. Hosch, "Google," *Britannica Money*, Jul. 2024.

[184] A. Tyson, B. Kennedy, and C. Funk, "Gen Z, Millennials Stand Out for Climate Change Activism, Social Media Engagement With Issue," Pew Research Center, Tech. Rep., May 2021.

[185] S. Abendschön and G. García-Albacete, "It's a man's (online) world. Personality traits and the gender gap in online political discussion," *Information, Communication & Society*, vol. 24, no. 14, pp. 2054–2074, Oct. 2021.

[186] L. Vochocová, V. Štětka, and J. Mazák, "Good girls don't comment on politics? Gendered character of online political participation in the Czech Republic," *Information, Communication & Society*, vol. 19, no. 10, pp. 1321–1339, Oct. 2016, Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2015.1088881.

[187] K. Koc-Michalska, A. Schiffrin, A. Lopez, S. Boulianne, and B. Bimber, "From Online Political Posting to Mansplaining: The Gender Gap and Social Media in Political Discussion," *Social Science Computer Review*, vol. 39, no. 2, pp. 197–210, Apr. 2021, Publisher: SAGE Publications Inc.

[188] K. Solovev and N. Pröllochs, "Hate Speech in the Political Discourse on Social Media: Disparities Across Parties, Gender, and Ethnicity," in *Proceedings of the ACM Web Conference 2022*, Virtual Event, Lyon France: ACM, Apr. 2022, pp. 3656–3661, ISBN: 978-1-4503-9096-5.

[189] *Amdt26.1 Overview of Twenty-Sixth Amendment, Reduction of Voting Age*.

[190] B. Kim and J. Hoewe, "Developing contemporary factors of political participation," *The Social Science Journal*, vol. 60, no. 4, pp. 862–876, Oct. 2023, Publisher: Routledge _eprint: https://doi.org/10.1080/03623319.2020.1782641.

[191] F. Zhong, P. Li, and J. Xi, "A Survey on Online Political Participation, Social Capital, and Well-Being in Social Media Users—Based on the Second Phase of the Third (2019) TCS Taiwan Communication Survey Database," *Frontiers in Psychology*, vol. 12, 2022.

[192] Y. Sairambay, "Reconceptualising political participation," *Human Affairs*, vol. 30, no. 1, pp. 120–127, Jan. 2020, Publisher: De Gruyter.

[193] J. W. v. Deth, "What Is Political Participation?" In *Oxford Research Encyclopedia of Politics*, Oxford University Press, Nov. 2016, ISBN: 978-0-19-022863-7.

[194] Y.-H. Lee and G. Hsieh, "Does slacktivism hurt activism?: The effects of moral balancing and consistency in online activism," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Paris France: ACM, Apr. 2013, pp. 811–820, ISBN: 978-1-4503-1899-0.

[195] S. U. Noble, S. T. Roberts, M. Bui, A. Brock, and O. Snow, "Structural Racism in Tech: Social Media Platforms, Algorithmic Bias, and Racist Tech," in *Handbook of Children and Screens: Digital Media, Development, and Well-Being from Birth Through Adolescence*, Springer Nature Switzerland Cham, 2024, pp. 269–274.

[196] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, Jan. 2006, Publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa.

[197] J. Saldana, *The Coding Manual for Qualitative Researchers*. SAGE, Nov. 2015, Google-Books-ID: jh1iCgAAQBAJ, ISBN: 978-1-4739-4358-2.

[198] S. Timmermans and I. Tavory, "Theory Construction in Qualitative Research: From Grounded Theory to Abductive Analysis," *Sociological Theory*, vol. 30, no. 3, pp. 167–186, Sep. 2012, Publisher: SAGE Publications Inc.

[199] J. Burrell, Z. Kahn, A. Jonas, and D. Griffin, "When Users Control the Algorithms: Values Expressed in Practices on Twitter," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, 138:1–138:20, Nov. 2019.

[200] S. Kim, C. Oh, W. I. Cho, D. Shin, B. Suh, and J. Lee, "Trkic G00gle: Why and How Users Game Translation Algorithms," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, 344:1–344:24, Oct. 2021.

[201] G. Benjamin, "#FuckTheAlgorithm: Algorithmic imaginaries and political resistance," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22, New York, NY, USA: Association for Computing Machinery, Jun. 2022, pp. 46–57, ISBN: 978-1-4503-9352-2.

[202] S. A. Munson and P. Resnick, "Presenting diverse political opinions: How and how much," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '10, New York, NY, USA: Association for Computing Machinery, Apr. 2010, pp. 1457–1466, ISBN: 978-1-60558-929-9.

[203] G. Veletsianos, S. Houlden, J. Hodson, and C. Gosse, "Women scholars' experiences with online harassment and abuse: Self-protection, resistance, acceptance, and self-blame," *New Media & Society*, vol. 20, no. 12, pp. 4689–4708, Dec. 2018, Publisher: SAGE Publications.

[204] Jillian York, "Elon Musk doesn't know what it takes to make a digital town square," *MIT Technology Review*, Oct. 2022.

[205] B. Chappell, "Violence erupts at UCLA as protests over Israel's war in Gaza escalate across the U.S.," *NPR*, May 2024.

[206] A. International, *Toxic Twitter - The Silencing Effect*, Mar. 2018.

[207] R. Quinn, "Tenured Jewish Professor Says She's Been Fired for Pro-Palestinian Speech," *Inside Higher Ed*, 2024.

[208] C. Cohen and J. Kahne, "Participatory Politics: New Media and Youth Political Action," *Youth and Participatory Politics Research Network*, Jun. 2012.

[209] D. Rotman *et al.*, "From slacktivism to activism: Participatory culture in the age of social media," in *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '11, Vancouver, BC, Canada: Association for Computing Machinery, 2011, pp. 819–822, ISBN: 9781450302685.

[210] E. Simpson, "Integrated & Alone: The Use of Hashtags in Twitter Social Activism," in *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, ser. CSCW '18 Companion, New York, NY, USA:

Association for Computing Machinery, Oct. 2018, pp. 237–240, ISBN: 978-1-4503-6018-0.

[211] J. Hornsby, "Free speech and hate speech: Language and rights," *Normativita Fatti Valori*, Jan. 2003.

[212] N. Jankowicz, J. Hunchak, A. Pavliuc, C. Davies, S. Pierson, and Z. Kaufmann, "Malign Creativity: How Gender, Sex, and Lies Are Weaponized Against Women Online," The Wilson Center, Tech. Rep., 2021.

[213] "#ToxicTwitter: Violence and Abuse Against Women Online," Amnesty International, Tech. Rep., Mar. 2018.

[214] S. Neuman, "1 year after Hamas attacked Israel, the conflict grows more dangerous than ever," *NPR*, Oct. 2024.

[215] R. Shuy, *Language Crimes: The Use and Abuse of Language Evidence in the Courtroom*. John Wiley & Sons, Jun. 1996, Google-Books-ID: _BaBEAAAQBAJ, ISBN: 978-0-631-20153-3.

[216] T. Christensen, "Indirect threats as an illegal speech act," in *The Sign of the V. : Papers in Honour of Sten Vikner*, Aarhus Universitet, Jan. 2019, ISBN: 978-87-7507-461-7.

[217] M. T. Center, *How review teams work — Transparency Center*.

[218] M. J. Abrams, "Uncovering the Genre of Threatening Texts: A Multilayered Corpus Study," Accepted: 2019-07-05T18:28:51Z Publication Title: Georgetown University-Graduate School of Arts & Sciences, thesis, Georgetown University, 2019.

[219] A. Marsters, "When Hate Speech Leads to Hateful Actions: A Corpus and Discourse Analytic Approach To Linguistic Threat Assessment of Hate Speech," Ph.D. dissertation, Georgetown University, 2019.

[220] X. H. Center, *Hateful Conduct*, Apr. 2023.

[221] X. H. Center, *How X handles abuse and harassment — X Help*.

[222] *Counterman v. Colorado 600 U.S 66*, 2023.

[223] M. A. Franks, *How Stalking Became Free Speech: Counterman v. Colorado and the Supreme Court's Continuing War on Women* -, Jul. 2023.

[224] Morris and Dewitt, *Online Discourse and Cyberbullying: What Constitutes a Credible Threat?* Section: Supreme Court, Aug. 2023.

[225] *Elonis v. United States, 575 U.S. ___*, 2015.

[226] A. Johnston, *Why Today's Elonis Decision is a Victory in the Fight Against Online Harassment*, Jun. 2015.

[227] P. Casula, A. Anupam, and N. Parvin, ""We found no violation!": Twitter's Violent Threats Policy and Toxicity in Online Discourse," in *C&T '21: Proceedings of the 10th International Conference on Communities & Technologies - Wicked Problems in the Age of Tech*, ser. C&T '21, New York, NY, USA: Association for Computing Machinery, Jun. 2021, pp. 151–159, ISBN: 978-1-4503-9056-9.

[228] T. A. Gales, "Ideologies of Violence: A Corpus and Discourse Analytic Approach to Stance in Threatening Communications," ISBN: 9781124222974 Publication Title: ProQuest LLC ERIC Number: ED520035, Ph.D. dissertation, ProQuest LLC, 2010.

[229] T. Gales, "Threatening Stances : A corpus analysis of realized vs. non-realized threats," *Language and Law / Linguagem e Direito*, vol. 2, no. 2, 2015, Number: 2.

[230] B. Fraser, "Threatening revisited," *Forensic Linguistics-the International Journal of Speech Language and The Law - FORENSIC LINGUIST*, vol. 5, pp. 159–173, Dec. 1998.

[231] T. I. Emerson, "Toward a General Theory of the First Amendment," *The Yale Law Journal*, vol. 72, no. 5, p. 877, Apr. 1963.

[232] F. Askin, "Chilling Effect," *The Free Speech Center*, 2009.

[233] S. Gorissen, "Weathering and weaponizing the #TwitterPurge: Digital content moderation and the dimensions of deplatforming," *Communication and Democracy*, vol. 58, no. 1, pp. 1–26, Jan. 2024.

[234] C. Meisner, "The weaponization of platform governance: Mass reporting and algorithmic punishments in the creator economy," *Policy & Internet*, vol. 15, no. 4, pp. 466–477, 2023.

[235] A. Liptak, "Twitter says it 'made a mistake' for not removing threatening tweets from Florida bomb suspect," *The Verge*, Oct. 2018.

[236] "DSA Transparency Report," X, Tech. Rep., Sep. 2024.

[237] S. T. Roberts, "Digital detritus: 'Error' and the logic of opacity in social media content moderation," *First Monday*, Mar. 2018.

[238] V. E. L. Vinter, "You Call me a Bitch Like It's a Bad Thing : A Study into the Current Use and Semantic Properties of the Noun Bitch," Ph.D. dissertation, Mälardalen University, School of Education, Culture and Communication, 2017.

[239] S. Sobieraj, "Bitch, slut, skank, cunt: Patterned resistance to women's visibility in digital publics," *Information, Communication & Society*, vol. 21, no. 11, pp. 1700–1714, Nov. 2018, Publisher: Routledge.

[240] K. Riemer and S. Peter, "Algorithmic audiencing: Why we need to rethink free speech on social media," *Journal of Information Technology*, vol. 36, no. 4, pp. 409–426, Dec. 2021, Publisher: SAGE Publications Ltd.

[241] Cybersecurity and I. S. Agency, *Election Security Rumor vs. Reality — CISA*.

[242] Cybersecurity and I. S. Agency, *Foreign Influence Operations and Disinformation — Cybersecurity and Infrastructure Security Agency CISA*.

[243] Michael Blanding, *An Argument for Free Speech, the "Lifeblood of Democracy"*, Feb. 2024.

[244] M. T. Center, *Inauthentic Behavior — Transparency Center*.

[245] T. Huang, "Free Speech Capability," *Harvard Human Rights Journal*, Jun. 2024.

[246] M. H. Redish, "Value of Free Speech," *University of Pennsylvania Law Review*, vol. 130, 1982.

[247] T. I. Emerson, *The System Of Freedom Of Expression*. Random House, 1970.

[248] C. M. Koggel, A. Harbin, and J. J. Llewellyn, "Feminist relational theory," *Journal of Global Ethics*, vol. 18, no. 1, pp. 1–14, Jan. 2022, Publisher: Routledge _eprint: https://doi.org/10.1080/17449626.2022.2073702.

[249] S. H. Williams, "Democracy, Freedom Of Speech, And Feminist Theory: A Response To Post And Weinstein," *Virginia Law Review*, vol. 97, 2011.

[250] S. H. Williams, "Feminist Theory and Freedom of Speech," *Indiana Law Journal*, vol. 84, no. 3, 2009.

[251] J. Chan, "Understanding Free Speech as a Two-Way Right," *Political Philosophy*, vol. 1, no. 1, Jun. 2024, Number: 1 Publisher: Open Library of Humanities.

[252] M. Bonotti and J. Seglow, "Freedom of speech: A relational defence," *Philosophy & Social Criticism*, vol. 48, no. 4, pp. 515–529, May 2022.

[253] B. Weiser, "Mail Bomb Suspect Accused of Targeting Clinton, Obama and Other Democrats to Plead Guilty," *The New York Times*, Mar. 2019.

[254] D. O'Sullivan, "Bomb suspect threatened people on Twitter, and Twitter didn't act," *CNN Business*, Oct. 2018.

[255] M. A. Chang *et al.*, "Co-design Partners as Transformative Learners: Imagining Ideal Technology for Schools by Centering Speculative Relationships," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI '24, New York, NY, USA: Association for Computing Machinery, May 2024, pp. 1–15, ISBN: 9798400703300.

[256] C. Harrington and T. R. Dillahunt, "Eliciting Tech Futures Among Black Young Adults: A Case Study of Remote Speculative Co-Design," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI '21, New York, NY, USA: Association for Computing Machinery, May 2021, pp. 1–15, ISBN: 978-1-4503-8096-6.

[257] M. Blythe and E. Encinas, "The Co-ordinates of Design Fiction: Extrapolation, Irony, Ambiguity and Magic," in *Proceedings of the 2016 ACM International Conference on Supporting Group Work*, ser. GROUP '16, New York, NY, USA: Association for Computing Machinery, Nov. 2016, pp. 345–354, ISBN: 978-1-4503-4276-6.

[258] R. Y. Wong, E. Van Wyk, and J. Pierce, "Real-Fictional Entanglements: Using Science Fiction and Design Fiction to Interrogate Sensing Technologies," in *Proceedings of the 2017 Conference on Designing Interactive Systems*, ser. DIS '17, New York, NY, USA: Association for Computing Machinery, Jun. 2017, pp. 567–579, ISBN: 978-1-4503-4922-2.

[259] M. Blythe, "Research through design fiction: Narrative in real and imaginary abstracts," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '14, New York, NY, USA: Association for Computing Machinery, Apr. 2014, pp. 703–712, ISBN: 978-1-4503-2473-1.

[260] S. Klassen and C. Fiesler, "The Stoop: Speculation on Positive Futures of Black Digital Spaces," *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. GROUP, pp. 1–24, Jan. 2023.

[261] M. Webster, *Definition of AGORA*, Jan. 2025.

[262] N. Doderović, "Agora on the Internet: Can X's Spaces be interpreted as a Digital Platform for Democracy?" *Media Studies and Applied Ethics*, vol. 5, no. 2, pp. 47–61, Dec. 2024, Number: 2.

[263] J. J. Mark, "Agora," *World History Encyclopedia*, May 2021.

[264] R. Lewis, *Forum — Public Gathering Place, Marketplace & Political Hub*, Dec. 2024.

[265] Reddit, *Am I the Asshole?*

[266] d. boyd danah, "Social Network Sites as Networked Publics: Affordances, Dynamics, and Implications," in *A Networked Self*, Num Pages: 20, Routledge, 2010, ISBN: 978-0-203-87652-7.

[267] C. M. Rodrigo, *Jordan confronts tech CEOs over claims of anti-conservative bias*, Text, Jul. 2020.

[268] O. Donie and E. Alaa, *Twitter bans posts wishing for Trump death. The Squad wonders where that policy was for them*, Oct. 2020.

[269] L. D. Meco, "#ShePersisted: Women, Politics & Power in the New Media World.," The Wilson Center, Tech. Rep., 2019.

[270] N. McDonald and A. Forte, "Powerful Privacy Norms in Social Network Discourse," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–27, Oct. 2021.

[271] J. BRENNER, "FEMINIST POLITICAL DISCOURSES:: Radical Versus Liberal Approaches to the Feminization of Poverty and Comparable Worth," *Gender & Society*, vol. 1, no. 4, pp. 447–465, Dec. 1987, Publisher: SAGE Publications Inc.

[272] M. J. Matsuda, C. R. L. Iii, R. Delgado, and K. W. Crenshaw, *Words That Wound: Critical Race Theory, Assaultive Speech, And The First Amendment*. Routledge, Mar. 2018, ISBN: 978-0-429-97149-5.

[273] E. Goldberg, "Free Speech Consequentialism," *COLUMBIA LAW REVIEW*, vol. 116, no. 3,

[274] J. Wilson, "Political Discourse," in *The Handbook of Discourse Analysis*, Section: 20 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470753460.ch21, John Wiley & Sons, Ltd, 2005, pp. 398–415, ISBN: 978-0-470-75346-0.

[275] J. M. Balkin, "Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society," *NEW YORK UNIVERSITY LAW REVIEW*, vol. 79,

[276] R. Y. Wong, J. C. Valdez, A. Alexander, A. Chiang, O. Quesada, and J. Pierce, "Broadening Privacy and Surveillance: Eliciting Interconnected Values with a Sce-

narios Workbook on Smart Home Cameras," in *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, ser. DIS '23, New York, NY, USA: Association for Computing Machinery, Jul. 2023, pp. 1093–1113, ISBN: 978-1-4503-9893-0.

[277] S. Bardzell, J. Bardzell, J. Forlizzi, J. Zimmerman, and J. Antanitis, "Critical design and critical theory: The challenge of designing for provocation," in *Proceedings of the Designing Interactive Systems Conference*, ser. DIS '12, Newcastle Upon Tyne, United Kingdom: Association for Computing Machinery, 2012, pp. 288–297, ISBN: 9781450312103.

[278] G. R. Hayes, "The relationship of action research to human-computer interaction," *ACM Trans. Comput.-Hum. Interact.*, vol. 18, no. 3, Aug. 2011.

[279] M. Ratto, "Critical making: Conceptual and material studies in technology and social life," *The Information Society*, vol. 27, no. 4, pp. 252–260, Jul. 2011.

[280] Steven Overly, *John Samples discusses CEO Facebook Mark Zuckerberg removing the fact-checking and moving towards community notes on the POLITICO Tech podcast*.

[281] R. Rogers and S. Niederer, "The politics of social media manipulation," in *The Politics of Social Media Manipulation*, R. Rogers and S. Niederer, Eds., Amsterdam University Press, 2020, pp. 19–70.

[282] S. O. Lam, T. M. Lenard, and J. Samples, *John Samples on The Oversight Board and Content Moderation - Publications*, Jan. 2025.

[283] T. Verge, "The fallout of Meta's content moderation overhaul," *The Verge*, Jan. 2025.

[284] D. Kayyali, *Meta's Content Moderation Changes are Going to Have a Real World Impact. It's Not Going to be Good. — TechPolicy.Press*, Jan. 2025.

[285] K. Conger and D. Alba, "Twitter Comes Under Attack From Trump's Supporters," *The New York Times*, May 2020.

[286] Justin Vallejo, "ACLU warns of 'unchecked power' of big tech for banning Trump," *The Independent*, Jan. 2021, Section: News.

[287] J. Nicas and D. Alba, "Amazon, Apple and Google Cut Off Parler, an App That Drew Trump Supporters," *The New York Times*, Jan. 2021.

VITA